ARCADIS Design & Consultancy for natural and built assets

Glenn Springs Holdings, Inc.
A subsidiary of Occidental Petroleum Corporation

BASELINE ECOLOGICAL RISK ASSESSMENT

APPENDIX A: SEDIMENT QUALITY TRIAD ASSESSMENT

Newark Bay Study Area

Newark, New Jersey

~~November 2019~~May 2020

CONTENTS

TABLES

FIGURES

ATTACHMENTS

ACRONYMS AND ABBREVIATIONS

^{o}C	degrees Celsius
α	intercept parameter
%	percent
π	probability of mortality
A. abdita	*Ampelisca abdita*
AIC	Akaike's Information Criteria
ASTM	American Society for Testing and Materials
AVS/SEM	acid volatile sulfide/simultaneously extracted metal
B-IBI	benthic index of biotic integrity
β_i	slope parameter of the *ith* predictor variable
BERA	Baseline Ecological Risk Assessment
BIC	benthic invertebrate community
C	threshold mortality
CDF	confined disposal facility
cm	centimeters
CPG	Cooperating Parties Group
CV	coefficient of variation
DDx	dichlorodiphenyldichloroethane, dichlorodiphenyldichloroethylene, and dichlorodiphenyltrichloroethane
EA	EA Engineering, Science and Technology, Inc.
ERL	effects range-low
ERM	effects range-median
F	a cumulative probability distribution function
f	fraction
FCV	final chronic value
FOD	Frequency of Detection
GLM	General Linear Model
GSH	Glenn Springs Holdings, Inc.
HMW	high molecular weight

ln(L)	log likelihood
K_{PS}	polymer-water partition constant
L	liter
L. plumulosus	*Leptocheirus plumulosus*
LMW	low molecular weight
LOE	line of evidence
LPIL	lowest practicable identification level
LPRRP	Lower Passaic River Restoration Project
LPRSA	Lower Passaic River Study Area
m^2	square meter
MDD	minimum detectable difference
mg/kg	milligrams per kilogram
mg/L	milligrams per liter
mg/organism/day	milligrams per organism per day
NBSA	Newark Bay Study Area
ng/kg	nanograms per kilogram
NJ	New Jersey
NJDEP	New Jersey Department of Environmental Protection
NMDS	non-metric multidimensional scaling
NOAA	National Oceanic and Atmospheric Administration
Normandeau	Normandeau Associates, Inc.
NY	New York
OC	organic carbon
ORP	oxidation-reduction potential
p	the number of parameters in the model, including the intercept
PAH	polycyclic aromatic hydrocarbon
PCA	principal component analysis
PC	principal component
PC1	first principal component
PC2	second principal component
PC3	third principal component

PCB	polychlorinated biphenyl
PE	polyethylene
PEL	probable effect level
pg/L	picograms per liter
POM	polyoxymethylene
ppt	parts per thousand
PRC	performance reference compound
R	response
R^2	coefficient of determination
REMAP	Regional Environmental Monitoring Assessment Program
QAPP	quality assurance project plan
SQuiRT	Screening Quick Reference Tables
SQT	Sediment Quality Triad
SumTU	sum of toxic units
TCDD	tetrachlorodibenzo-*p*-dioxin
TEL	threshold effect level
TOC	total organic carbon
U.S.	United States
UMBC	University of Maryland Baltimore County
USACE	United States Army Corps of Engineers
USEPA	United States Environmental Protection Agency
WOE	weight of evidence
WQC	water quality criteria
x_i	*ith* predictor variable
µg/kg	micrograms per kilogram

1 INTRODUCTION

This appendix provides the results of the comprehensive analyses of sediment, porewater, toxicity, and benthic invertebrate community (BIC) data collected in the Newark Bay Study Area (NBSA) in support of the assessment of risks to benthic invertebrates posed by chemical stressors in sediments.

The Sediment Quality Triad (SQT) and Porewater Sampling Program was conducted in 2015, as described in the NBSA SQT and Porewater Sampling and Analysis, Quality Assurance Project Plan (QAPP) (Tierra 2015); the NBSA SQT and Porewater Field Report (Glenn Springs Holdings, Inc. [GSH] 2017a); and the NBSA SQT and Porewater Data Report (GSH 2017b). SQT sampling was conducted at 30 locations in the NBSA (Figure A-1). Surface sediment samples at each of the 30 locations, herein referred to as the SQT stations, were collected from the top 6 inches using grab sampling techniques, and the analysis consisted of the following components: 1) surface sediment chemistry[1]; 2) porewater chemistry; 3) toxicity testing (10-day and 28-day *Leptocheirus plumulosus*); and 4) BIC taxa enumeration.

The SQT program was designed to support an effects-based, weight of evidence (WOE) assessment to evaluate risks to benthic invertebrate communities using three primary lines of evidence (LOEs): 1) sediment chemistry, 2) sediment toxicity, and 3) BIC. The approach is widely utilized by regulatory agencies, including the United States Environmental Protection Agency (USEPA) Region 2 and its partner agencies[2], as the most appropriate process currently available for assessing potential risks to benthic invertebrates. It is the approach that is being used to assess risks to benthic invertebrates for the Lower Passaic River Restoration Project (LPRRP). For consistency purposes and data comparability, this approach is also used for the NBSA.

The SQT method incorporates a combination of both quantitative and qualitative analyses to identify potential cause-and-effect relationships between two or more of the components. The approach provides information needed to identify degraded conditions and to assess the impacts of pollution-induced effects on benthic invertebrates (Tierra 2015). The specific technical approach for the SQT was developed with USEPA in an iterative and collaborative manner between 2015 and 2018.

Section 2 of this appendix describes the SQT evaluation; an objective method of comparing sediment/porewater chemistry concentrations; sediment toxicity endpoints; and BIC metrics to screening, control, or reference values with the objective of categorizing each SQT sampling station with respect to the degree of potential impact to benthic invertebrates. Section 3 of this appendix provides a quantitative assessment of effects, incorporating bivariate and multivariate statistical methods to evaluate the association of sediment and porewater chemistry with effects measures (i.e., toxicity endpoints and BIC metrics). Section 4 presents a summary of LOEs for the potential risk to the BIC in the NBSA posed by sediment contamination.

[1] During the SQT program, additional samples were collected at 13 locations (i.e., for sediment chemistry only) for a total of 43 sediment chemistry samples to support the baseline human health risk assessment.
[2] The partner agencies included the United States Army Corps of Engineers (USACE), the New Jersey Department of Environmental Protection (NJDEP), the New Jersey Department of Transportation, the National Oceanic and Atmospheric Administration (NOAA), and the U.S. Fish and Wildlife Service.

2 SEDIMENT QUALITY TRIAD EVALUATION

This evaluation was used to classify each of the 30 SQT stations with respect to the degree of potential impact. Three categories (i.e., the triad) were evaluated: BIC composition, sediment toxicity, and sediment/porewater chemistry. Each category has equal weighting in the evaluation and was assigned a score ranging from 0 (unimpacted) to 1 (most impacted). The scores for each category are summed (maximum possible score = 3), and the final score is used to categorize each station with respect to the degree of impact. Consistent with the Lower Passaic River Study Area (LPRSA) Baseline Ecological Risk Assessment (BERA) (Windward 2019) and as agreed upon with USEPA, the final scores were assigned impact categories as follows: no impact = <0.75; low impact = >0.75 and <1.5; medium impact = >1.5 and <2.25; and high impact = >2.25.

2.1 Benthic Invertebrate Line of Evidence

2.1.1 Sampling Methods

Benthic invertebrate sampling was conducted according to the methods described in the SQT and Porewater Sampling and Analysis QAPP (Tierra 2015), the SQT and Porewater Field Report (GSH 2017a), and the SQT Data Report (GSH 2017b). Normandeau Associates, Inc. (Normandeau) of Bedford, New Hampshire, conducted BIC analyses on samples collected from the 30 SQT stations. Three replicate samples were collected from each sampling station using either a petite Ponar dredge (area of 0.023 square meters [m^2]) or a standard Ponar dredge (area of 0.052 m^2). The samples were immediately sieved through a standard number 30 (600 microns) sieve on the field vessel, and the retained invertebrates were preserved with isopropyl alcohol (90%). Preserved samples were transported to Normandeau's laboratory for analysis at the end of the sampling program. Organisms were identified to the genus/species (lowest practicable) phylogenetic endpoint using dissection and compound microscopes. In addition to the taxonomic analysis, Normandeau also measured the dry-weight (shell-free) biomass of the 30 benthic invertebrate samples. Biomass for each sample was determined using Standard Methods for the Examination of Water and Wastewater: Method 209 A – Total Residue Dried at 103 to 105^{o}C (American Public Health Association, American Water Works Association, and Water Pollution Control Federation 1980). All organisms were included in each sample weight. Shelled organisms (i.e., mollusks) had the tissue removed from the shell, and the shells were not included in the dry weight. Samples were weighed to 0.0001 grams using an analytical balance.

The taxa identified and the number of organisms per m^2 of each taxon at each SQT station are provided in Table A-1.

2.1.2 Benthic Invertebrate Community Metrics

Taxa in the BIC samples were identified to the lowest practicable identification level (LPIL) (Table A-1) in a manner consistent with other surveys performed in the New York (NY)/New Jersey (NJ) Harbor Estuary (Weisberg et al. 1998). The following six BIC metrics were calculated for each sample:

- Abundance (i.e., density): the number of organisms per m^2

- Number of taxa (i.e., number of discrete LPIL taxa)
- Shannon-Weiner Diversity Index (Shannon 1948)
- Pielou's Evenness Index (Pielou 1966)
- Swartz's Dominance Index (Swartz et al. 1985)
- NY/NJ Benthic Index of Biotic Integrity (B-IBI) (Weisberg et al. 1998)

Abundance is reported as density (organisms per m^2), which is calculated by dividing the number of organisms counted by the total sampling area of the sampling device. The number of taxa is the total number of species observed in the sample.

The Shannon-Wiener Diversity Index (Shannon 1948) is a measure of the species diversity of a sample and considers species richness and the evenness of each species within the community. The number of species (or higher order taxa, if not identified to species) and the number of individuals within each taxa are combined to calculate the Diversity Index. The index value ranges from 0 to the natural logarithm of the number of taxa in the sample. The greater the value, the greater the sample diversity and the more evenly the organisms are distributed among the taxa present. It is calculated as follows:

$$H' = -\Sigma\, p_i \ln p_i$$

Where:

H' = the Shannon-Wiener Diversity Index

ln = the natural logarithm

Σ = summation

p_i = the number of individuals within the taxon *i* divided by the total number of individuals present in the entire sample.

Pielou's Evenness Index (Pielou 1966) is related to the Shannon-Wiener Diversity Index (Shannon 1948) in that it also examines the distribution of individuals among taxa relative to an idealized distribution. Pielou's Evenness Index ranges from 0.0 to 1.0, with the higher values indicating a more even distribution of organisms among the species present in the sample. It is calculated as follows:

$$J = H' / H(max)$$

Where:

J = Pielou's Evenness Index

H' = the Shannon-Wiener Diversity Index

H(max) = the theoretical maximum value for H' if all species in the sample were equitably distributed (i.e., natural logarithm S, where S is the total number of taxa in a sample).

The B-IBI is calculated using a method derived for the saline portions of the NY/NJ Harbor Estuary (Weisberg et al. 1998). This method incorporates abundance, number of taxa, percent pollution tolerant, percent pollution sensitive, and biomass into one index using comparisons to unimpacted reference sites

(Table A-2). Threshold values of the B-IBI for the classification of impacts are as follows: 1 to <2 = impacted, 2 to <3 = slightly impacted, and 3 to 5 = unimpacted (Weisberg et al. 1998). Figure A-2 shows the B-IBI scores for the SQT stations in the NBSA.

2.1.3 Comparison to Reference Data

As agreed upon with USEPA, the reference site chosen for this evaluation is the Jamaica Bay Estuary, New York, which is physically similar to the open water habitats of the NBSA. A total of 150 SQT samples (i.e., sediment samples with co-located chemistry, toxicity [i.e.,10-day *Ampelisca abdita* bioassay], and BIC data) were collected from Jamaica Bay over of period of years (1993 through 2013) by USEPA under the Regional Environmental Monitoring Assessment Program (REMAP) and by other agencies. A screening process was conducted to select a subset of these 150 samples to represent reference conditions for the BIC. The Jamaica Bay sediment chemistry data were compared to effects range-low (ERL) and effects range-median (ERM) values (Long et al. 1995). Acceptable Jamaica Bay reference locations had three or fewer exceedances of ERLs and no exceedances of ERMs across all chemicals for which those sediment guidelines were available. In addition to meeting the chemical criteria, acceptable reference locations in Jamaica Bay were required to meet sediment toxicity criteria. Specifically, *A. abdita* survival results at Jamaica Bay reference locations were required to be ≥ 80% of the respective negative control response. For data collected prior to 2008, the screening assessment was conducted by the Cooperating Parties Group (CPG) for use as reference data for the LPRSA (Windward 2019) and resulted in 35 acceptable samples. The screening of data collected in 2008 and 2013 under REMAP was conducted by GSH for use in this SQT assessment and included an additional 24 samples. The combined dataset includes a total of 59 SQT samples from Jamaica Bay that are suitable for use as reference data in a reference envelope evaluation. Attachment A-1 provides the final Jamaica Bay BIC and toxicity reference data set (Table A-1-1); the 2008/2013 REMAP sediment chemistry screening results (Table A-1-2); the 2008/2013 REMAP toxicity screening results (Table A-1-3); the 2008/2013 REMAP density and biomass data (Table A-1-4); the 2008/2013 REMAP sediment chemistry statistics (Table A-1-5); and the sediment chemistry data for the final reference data set (Table A-1-6). A comparison of grain size ~~and~~, organic carbon, bottom dissolved oxygen and bottom salinity between Jamaica Bay and Newark Bay is also shown in Attachment A-1 (Table A-1-7). The mean percent silt and clay is 36% in Jamaica Bay and 56% in Newark Bay. The mean total organic carbon (TOC) is 2.5% in Jamaica Bay and 2.4 % in Newark Bay. Bottom dissolved oxygen is similar in the two estuaries with a mean range of 5 to 5.25 milligrams per liter (mg/L). Jamaica Bay is generally more saline (mean = 29 parts per thousand [ppt]) than Newark Bay (mean = 21 ppt). Probability plots comparing sediment chemistry data for the Jamaica Bay reference data set and the 30 SQT samples for selected COCs are also shown in Attachment A-1 (Figures A-1-1, A-1-2 and A-1-3).

Figures A-3 and A-4 present box plots of Jamaica Bay and NBSA BIC metrics by geographic zone along with the results of non-parametric (i.e., Kruskal-Wallis) hypothesis tests. Mean abundance in Jamaica Bay is greater than that of the NBSA; however, because abundance can either decrease or increase with degradation (Weisberg et al. 1998), inferences regarding comparability are uncertain for this metric and, therefore, no hypothesis testing was conducted. The mean number of taxa in Jamaica Bay (19) is lower than that of the NBSA (21 to 26), but the difference is not statistically significant (Figure A-3). For other metrics, including the Shannon Wiener Diversity Index, Pielou's Evenness Index, and Swartz's Dominance Index, the Jamaica Bay reference means are significantly lower than the NBSA means

(Figure A-4). This comparison indicates that the NBSA BIC are not impaired in comparison to Jamaica Bay. It should be noted that the SQT samples were collected at a depth of 0 to 15 centimeters (cm) while those in Jamaica Bay were collected at a depth of 0 to 2 cm, which adds some degree of uncertainty to the BIC comparison.

2.1.4 Scoring

Based on the BIC composition, each SQT station was assigned a score ranging from 0 (unimpacted) to 1 (most impacted) (Table A-3) based on a reference envelope approach. This approach involves the comparison of BIC metrics measured at the potentially impacted SQT stations to lower (or upper) percentiles of the distribution of metric values measured at the reference site. Four of the five BIC metrics (number of taxa, Shannon-Weiner Diversity Index, Pielou's Evenness Index, and Swartz's Dominance Index) are expected to decline in value as a result of physical and/or chemical degradation. Therefore, these metrics were compared to the 5th percentile of the Jamaica Bay reference distribution. BIC metrics that are equal to or greater than the 5th percentile of the reference distribution are considered comparable to reference. Abundance is a metric that can either decline or increase as a result of degradation (Weisberg et al. 1998). Therefore, abundance was compared to both the 5th and the 95th percentile of the Jamaica Bay reference distribution. Abundance equal to or greater than the 5th percentile and less than or equal to the 95th percentile of the Jamaica Bay reference distribution is considered comparable to reference. Metrics outside the reference envelope received a score of 1 (Table A-3). The reference percentiles were calculated using a non-parametric method based on order statistics. The equation for percentile calculation is given in Attachment A-1. For the B-IBI, a score of either 0, 0.5, or 1 was applied according to the impact categories described by the authors (Weisberg et al.1998) with respect to degree of impact (Table A-3). Each metric score was weighted equally (0.167), and the weights were multiplied by the metric score and summed such that the maximum possible score for the BIC is 1, which represents the greatest impact.

Table A-4 presents the metric values and SQT scores for each of the 30 stations. The Jamaica Bay reference envelope values are shown for each metric. For the five standardized metrics, only one value (abundance at Station 148) was outside the Jamaica Bay reference envelope. Based on the B-IBI values, 21 stations were classified as non-impacted and eight stations were classified as slightly impacted (three in the north, one in the central, and four in the south). Station 148, which is located along a bulkheaded portion of the southern shoreline of the NBSA (Figure A-1), was classified as impacted. The combined BIC SQT scores are very low for most stations, ranging from 0 to 0.08 indicating non-impacted BIC. The BIC at Station 148, with a score of 0.33, exhibited low abundance, low number of taxa, and low B-IBI score; and thus, indicated a potential impact to the BIC at that location.

2.2 Sediment Toxicity Line of Evidence

2.2.1 Sampling Methods

Sediment sampling for toxicity testing and testing methods were conducted, as described in the SQT and Porewater Sampling and Analysis QAPP (Tierra 2015), the SQT and Porewater Field Report (GSH 2017a), and the SQT and Porewater Data Report (GSH 2017b). EA Engineering, Science and Technology, Inc. (EA) conducted two toxicity tests on the sediment samples collected from the 30 SQT

stations using the amphipod *L. plumulosus*: 1) a 10-day exposure for survival and 2) a 28-day exposure for survival, growth, and reproduction. Toxicity testing was conducted using standard methods: for 10-day tests (USEPA 1994; American Society for Testing and Materials [ASTM] 2008) and for 28-day tests (USEPA 2001). Laboratory control samples were tested coincidentally with SQT samples using sediment from the Pretty Boy Reservoir in Maryland. The *L. plumulosus* were acquired from Chesapeake Cultures, located in Hayes, Virginia.

The 10-day toxicity test was conducted from October 30, 2015 through November 9, 2015. Test chambers (i.e., 1-liter beakers) were prepared using sediment samples from the 30 SQT stations and the control sample (five replicates per sample). Artificial seawater was used as the overlying water in each test chamber. After a 7-day stabilization period, the test organisms were loaded into each test chamber (20 organisms per replicate) for a 10-day exposure period. During the 10-day exposure period, the test chambers were not aerated, and the test organisms were not fed. The test chambers were visually inspected daily for abnormal organism behavior or lack of burrowing, and water quality measurements of temperature, pH, dissolved oxygen, oxidation-reduction potential (ORP), and salinity were recorded daily. At the end of the 10-day exposure period, the test organisms were retrieved from the test chambers, and the number of live organisms in each chamber was recorded. The mean percent survival of the five replicates from each station and control was calculated.

The 28-day toxicity test was conducted from October 22, 2015 through November 19, 2015; however, due to unacceptable control mortality (25%), the test was considered invalid and was re-run. A second, successful test was conducted from November 25, 2015 through December 23, 2015. Test chambers were prepared using sediment samples from the 30 SQT stations and the control sample (five replicates per sample). Artificial seawater was used as the overlying water in each test chamber. After a 7-day stabilization period, the test organisms were loaded into each test chamber (20 organisms per replicate) for a 28-day exposure period. During the 28-day exposure period, the overlying water in each test chamber was gently aerated at a rate of 100 bubbles per minute throughout the exposure period, and the organisms were fed three times a week. The test chambers were visually inspected daily for abnormal organism behavior or lack of burrowing, and water quality measurements of temperature, pH, dissolved oxygen, ORP, and salinity were recorded daily. Ammonia measurements were also recorded during the 28-day exposure period at test initiation and termination. At the end of the 28-day exposure period, the test organisms were retrieved from the test chambers, and the number of live organisms (parent and offspring) in each chamber was recorded. The organisms were then dried and weighed. The mean of the five replicates from each station and control was calculated for the following endpoints: percent survival; growth, which is the change in mass over the 28-day period in milligrams per organism per day (mg/organism/day); and reproduction, which is the number of offspring per surviving adult.

Pursuant to USEPA guidance (USEPA 2001) for hypothesis testing of toxicity results, EA used a software program called ToxCalc™ V5.0 to process the endpoint data recorded for each replicate and to conduct hypothesis testing. With the ToxCalc™ V5.0 software, replicate data for each sample were compared to the control replicate data independently. The default test is a one-sided t-test ($\alpha = 0.05$). If the data did not meet the assumption of normality, then either a parametric test on transformed data (e.g., arcsin square root transform for survival data) or a non-parametric test (Wilcoxon Rank Sum) was conducted. For parametric tests, an equal variance test was also conducted to determine whether to report the p-value for an equal variance or unequal variance (Satterthwaite) t-test. The individual ToxCalc™ output files are attached to the SQT and Porewater Data Report (GSH 2017b). A summary of the bioassay data

is found in Attachment A-2 (Table A-2-1). The minimum detectable difference (MDD) from control is given for each endpoint. For 28-day survival, several samples that had mortality less than 80% of control mortality were not found to be statistically significantly different from control. This is because of the large variance seen in the replicate data. For example, the sample from station 138 had 46% survival (57% of control). But the replicate survival varied from 0 to 95% in this sample. This variability adds uncertainty to the interpretation of this endpoint. The individual replicate data are found in Table A-2-2.

2.2.2 Measured Endpoints

The replicate means for each endpoint at each SQT station and the control are provided in Table A-5. Shaded values indicate that the mean was statistically significantly different than the control mean. Bold values indicate the value is less than the control value minus the MDD. The 10-day survival was significantly different than the control at two stations (Stations 151 and 154), which were located on subtidal flats in the southwestern portion of the NBSA (Figure A-5). The 28-day survival was significantly different than the control at six stations, two in the north (Stations 158 and 160), two in the central (Stations 141 and 143), and two in the south (Stations 151 and 154) geographic areas of the NBSA (Figure A-5).

2.2.3 Comparison to Reference Data

Figure A-6 presents box plots that compare the 10-day amphipod survival in the NBSA by geographic area and Jamaica Bay. Ten-day tests for Jamaica Bay were conducted with the amphipod *A. abdita*. Although different organisms were used in the two estuaries, both *A. abdita* and *L. plumulosus* were found to be comparable in their ability to classify sediment samples as toxic in an inter-species and interlaboratory study conducted by Schlekat et al. (1995). A comparison of median survival was conducted using the non-parametric Kruskal-Wallis Test. Survival was first adjusted to control survival as follows: Control-adjusted survival = Observed survival/Control survival. The median control-adjusted survival in the NBSA was compared to the median control-adjusted survival in Jamaica Bay. The NBSA medians were statistically significantly greater than the Jamaica Bay reference median (Figure A-6). It is important to note that the 10-day survival in the Jamaica Bay reference dataset is biased high due to the screening process. All samples with observed survival of less than 80% of control were screened from the Jamaica Bay dataset regardless of whether the sample had elevated chemistry. Therefore, the actual median control survival in Jamaica Bay, based on all samples or even just those samples that would pass the chemistry screen, is lower than that based on the 59 samples that passed the chemistry and toxicity screen. Nonetheless, even when compared to this biased high-reference dataset, the 10-day survival in the NBSA is demonstrated to be comparable to Jamaica Bay.

2.2.4 Scoring

Based on the four toxicity endpoints, each SQT station was assigned a score ranging from 0 (unimpacted) to 1 (most impacted). The sediment toxicity data comprise survival and sublethal endpoints (i.e., growth and reproduction). Criteria for assigning toxicity are well established for the survival endpoints but not for the sublethal endpoints. Criteria used for REMAP (USEPA 2003a) defined samples as toxic if the 10-day amphipod (i.e., *A. abdita*) survival rate was less than 80% of the respective control survival and as highly toxic if the 10-day amphipod (i.e., *A. abdita*) survival rate was less than 60% of the

control survival. The 80% decision criterion is consistent with the criterion in the testing manual for discharge to United States (U.S.) waters (USEPA and USACE 1998) and was also suggested by Kennedy et al. (2009) for 28-day *L. plumulosus* toxicity tests. Therefore, these criteria were adopted to define the toxicity of sediments from the NBSA based on the survival endpoints. Where survival is >80% of control and/or not significantly different from control, a score of 0 was applied (Table A-6). Survival significantly different from control and between 60 and 80% of control received a score of 0.5. Survival significantly different from control and <60% of control received a score of 1 (Table A-6).

The sublethal endpoints are subject to considerable response variability compared to the survival endpoints both in test sediments and control sediments. Eickhoff et al. (2014) subjected five control sediments to 28-day tests with *L. plumulosus* and found little variation in survival among the five control samples, which ranged from 96 to 99%. The within batch-replicate variance was also low for survival, with a mean coefficient of variation (CV) of 5.6%. However, growth and reproduction endpoints were more variable both among control samples and within batch replicates. Mean growth in control samples ranged from 0.7 to 1.8 milligrams per organism (0.025 to 0.064 mg/organism/day), and the mean CV among replicates was 30%. Reproduction had a 10-fold range within control sediments (0.8 to 8.4 juveniles per surviving adult), and the mean CV among replicates was 79%. Kennedy et al. (2009) found significant variability in control sediment for growth (CV = 56%) and reproduction (CV = 63%), as well as significant variability in an interlaboratory comparison of the growth and reproduction endpoints. *L. plumulosus* reproduction is known to be highly dependent on grain size (McDonald et al. 2010). This fact was illustrated in Kennedy et al. (2009), where it was found that a clean reference sample, which was relatively coarse-grained, had a very low response for the reproductive endpoint compared to the control (i.e., less than 10% of the control response). Based on the typical criteria used to define toxicity for survival endpoints (i.e., less than 80% of the control response), the clean reference sample would be classified as toxic based on the reproductive endpoint.

Given the large variability that can occur among control sediments for these sublethal endpoints and because there are no established control acceptability criteria for these sublethal endpoints, a simple comparison to control sediment cannot be used to judge a sediment sample as toxic. Further, high variability makes default application of decision criteria (e.g., 20% reduction compared to control) inappropriate (McDonald et al. 2010). Given these considerations, the SQT scoring for the sublethal endpoints was adjusted as compared to the survival endpoints. Where sublethal response is >70% of control and/or not significantly different from control, a score of 0 was applied (Table A-6). Sublethal response significantly different from control and between 50 and 70% of control received a score of 0.5. Sublethal response significantly different from control and <50% of control received a score of 1 (Table A-6).

The NBSA toxicity data for sublethal endpoints was also subject to high variability (Attachment A-2). The CV for 28-day survival ranged from 4% to as high as 139% with variance generally increasing as mean survival decreases. ~~This variability can make statistical comparisons with control data challenging.~~ The average MDD at $\alpha = 0.05$ and $\beta = 0.20$ for this endpoint is ~~27~~33%. Therefore, some stations with survival less than 80% of control were not found to be statistically significantly different than control. The uncertainty of this with respect to SQT scoring is discussed in Section 4.1.

Each toxicity endpoint score was weighted equally (0.25), and the weights were multiplied by the score and summed such that the maximum possible score for the toxicity category of the triad is 1, which represents the greatest impact.

Table A-7 presents the scoring for the four sediment toxicity endpoints at each of the 30 SQT stations. Statistical significance is depicted by shading. Twenty-eight stations had 10-day survival comparable to the laboratory control survival (90%). Only two stations, Stations 151 and 154, located near the southwestern shoreline of the NBSA (Figure A-5), exhibited acute toxicity and received the maximum SQT score of 1. Thirteen stations received a score of 0 because all of the endpoints were comparable to the control based on the scoring criteria. Fifteen stations had scores between 0 and 1 based on a comparison of chronic endpoints to laboratory control. However, in total, only five stations had 28-day survival that was both less than 80% of control and significantly different than control (Stations 151, 154, 143, 158, and 160). The 28-day survival at Station 141 (67%), while significantly different than control, was greater than 80% of control (i.e., 83% of control) and, therefore, received a score of 0 for this endpoint. Alternative scoring methods are evaluated in the uncertainty analysis.

2.3 Sediment Chemistry Line of Evidence

2.3.1 Sampling Methods

Sediment and porewater sampling were conducted according to the methods described in the SQT and Porewater Sampling and Analysis QAPP (Tierra 2015), the SQT and Porewater Field Report (GSH 2017a), and the SQT and Porewater Data Report (GSH 2017b). Sediment samples were collected from the top 6 inches of sediment using grab sampling techniques (i.e., standard or petite Ponar dredge). Samples were sent to the appropriate laboratories for chemical analyses that included dioxins/furans, polychlorinated biphenyls (PCBs: congeners and Aroclors), pesticides, semivolatile organic compounds, volatile organic compounds, petroleum hydrocarbons, butyltins, herbicides, inorganic compounds, acid volatile sulfide/simultaneously extracted metals (AVS/SEM), organic carbon, and grain size.

Sediment samples were sent to the University of Maryland Baltimore County (UMBC) where an ex-situ (i.e., laboratory) porewater passive sampler study was conducted from September 23, 2015 through November 6, 2015. Hydrophobic organic compounds were measured using polyethylene (PE) (i.e., PCB congeners, pesticides, and polycyclic aromatic hydrocarbons [PAHs]) and polyoxymethylene (POM) (i.e., dioxins/furans) passive samplers. As described in the SQT and Porewater Sampling and Analysis QAPP (Tierra 2015), PE and POM samplers were placed in test chambers containing site sediment and site water, and the sediment slurry was allowed to mix via tumbling for 30 days. PE samplers were impregnated with a set of performance reference compounds (PRCs) that provided the assessment of the extent of equilibrium achieved during the contact period. Following exposure to sediment, the PE and POM samplers were cleaned and transferred to the analytical laboratories, where the target organic compounds were extracted in solvent (acetone/ hexane [1:1 by volume]) and the concentrations in final extracts were measured. Porewater concentrations were estimated using the laboratory-reported concentrations in the PE/POM extracts, as described by Ghosh et al. (2014). A simple mass balance calculation is used to convert extract concentrations to concentrations within the PE/POM. The porewater concentration is then estimated as the ratio of the PE or POM concentration and a literature-derived polymer-water partition constant (K_{PS}). The PRCs were used to assess the extent of equilibrium achieved

during the period that the samplers were in contact with the sediment. Some of the strongly hydrophobic compounds, in particular, the higher-chlorinated PCB congeners, did not achieve equilibrium during the period of contact, and a mathematical correction was performed to calculate the true equilibrium concentrations (Fernandez et al. 2009). The K_{PS} values used in the calculations were derived from literature: PCBs and parent PAHs (Ghosh et al. 2014); alkylated PAHs (Choi et al. 2013); isomers of dichlorodiphenyldichloroethane, dichlorodiphenyldichloroethylene, and dichlorodiphenyltrichloroethane (DDx) (Fernandez et al. 2014); other pesticides (Fernandez et al. 2012); and dioxins/furans (Cornelissen et al. 2008). For all other hydrophobic organic compounds, a correlation based on the octanol-water partition coefficient was used as described in Fernandez et al. (2012).

Dialysis (i.e., diffusion) samplers were used to measure metals and other inorganics (i.e., mercury, methylmercury, dissolved organic carbon, ammonia, and total sulfide) in porewater. Ex-situ laboratory equilibrium studies were conducted in 5-gallon buckets (one bucket for each sampling location). After the buckets were allowed to reconsolidate in the laboratory for 2 weeks, dialysis samplers that contained deionized water with matching site salinity were placed in each bucket. Multiple dialysis devices were placed within the top 12 inches of sediment in each bucket to allow the final collection of 1.5 liters (L) of dialyzed porewater. The dialysis samplers were allowed to equilibrate in the sediments for a period of 2 weeks. After removal, the water sample within the dialysis was prepared and shipped to the analytical laboratory for analysis. Because inorganic porewater concentrations were measured directly by the laboratory, no additional calculations were required.

Porewater analysis included: dioxins/furans, PCB congeners, pesticides, PAHs, metals, dissolved organic carbon, ammonia, and total sulfide.

2.3.2 Sediment Chemistry Screen

Sediment chemical concentrations were compared to screening values to assess the potential for impact to the BIC. Sediment screening values were preferentially based on those published either by USEPA (2005a) or Long et al. (1995). The USEPA (2005a) developed sediment screening values for a number of chemicals based on their predicted toxicity to amphipods. For each chemical – T20, T50, and T80 – values were developed that correspond to probabilities of toxicity of 20%, 50%, and 80%, respectively. Long et al. (1995) developed sediment screening values for a number of chemicals, ERL and ERM, that correspond to threshold and probable effect concentrations. Long et al. (2006) also present a method for assessing sediment quality using an ERM quotient, which is the sum of the chemical concentrations in sediment divided by the ERMs for 25 chemicals. An ERM quotient was calculated for each SQT sample. For chemicals with a frequency of detection (FOD) greater than 25% in surface sediment, for which T20/T50 and/or ERL/ERM values were not available, threshold and median screening values were selected based on threshold effect level (TEL) or probable effect level (PEL) values for marine sediment, which are published in the National Oceanic and Atmospheric Administration (NOAA) Screening Quick Reference Tables (SQuiRT) (Buchman 2008). Table A-8 provides a complete list of the sediment screening values used in this evaluation. A single threshold value is reported as the maximum of the T20 and ERL values, if available, or the TEL value, if not available. A single median value is reported as the maximum of the T50 or ERM values, if available, or the PEL, if otherwise. Table A-9 presents the number of exceedances by chemical for each of the screening values presented in Table A-8 in the 30 SQT samples. The threshold values for most chemicals were exceeded at most stations. Median values were

less frequently exceeded. The results of the screening for each station/analyte are provided in Attachment A-3 (Table A-3-1). Table A-10 provides a list of the chemicals for which no screening values were available.

2.3.3 Porewater Chemistry Screen

Porewater chemical concentrations were screened to assess the potential for impact to the BIC. In its derivation of equilibrium partitioning sediment benchmarks for the protection of benthic organisms, USEPA (2003b) conducted ancillary analyses that suggest that the sensitivity distribution of benthic and epibenthic organisms is not significantly different from that of pelagic organisms; therefore, water quality criteria (WQC) applied to interstitial waters should be fully protective of benthic organisms. As such, porewater screening values were preferentially based on the New Jersey acute and chronic WQC for aquatic life in saline water (NJDEP 2016). For detected compounds without WQC, acute and chronic screening values were selected form the NOAA SQuiRT tables (Buchman 2008) for marine surface waters. If marine guidelines were not available, freshwater guidelines were selected from the SQuiRT tables. Table A-11 provides a list of screening values for porewater. Table A-12 presents the number of exceedances by chemical for each of the screening values for the 30 porewater samples. Exceedances of chronic WQC were observed for PCBs and total DDx, but all other organic compound concentrations were below chronic screening values. Arsenic, copper, and zinc were infrequently detected in porewater, but where they were detected, they were often in exceedance of acute or chronic WQC. Of the five acute exceedances observed, three were for zinc, and one each was for arsenic and copper. The results of the screening for each station/analyte are provided in Attachment A-3 (Table A-3-2). Table A-13 provides a list of chemicals detected in porewater for which no screening values were available.

2.3.4 Other Screening Methods

The SQT data were also evaluated according to USEPA equilibrium sediment partitioning guidance for metal and PAH mixtures (USEPA 2003c, 2005b, 2017a). In accordance with the SQT and Porewater Sampling and Analysis QAPP (Tierra 2015), sediment samples were analyzed for AVS/SEM. Under equilibrium partitioning theory, AVS binds to divalent metals (i.e., cadmium, copper, lead, nickel, and zinc) such that they are not bioavailable in sediments and, therefore, are non-toxic to benthic organisms (USEPA 2005b). In addition, TOC contributes to the lack of bioavailability (USEPA 2005b). To evaluate the bioavailability of divalent metals, SEM concentrations are summed and the AVS concentrations are subtracted to give the value ∑SEM-AVS. If the value is negative, it is inferred that the metals are bound to the AVS in sediment and not bioavailable (USEPA 2005b). Further, ∑SEM-AVS is divided by the fraction (f) of organic carbon (OC) in the sample (∑SEM-AVS/f_{OC}). Based on toxicity experiments, when ∑SEM-AVS)/f_{OC} is less 130 micromoles per gram of OC, no toxicity is expected (USEPA 2005b). Porewater concentrations are evaluated directly by summing the concentration of each divalent metal (i.e., cadmium, copper, nickel, lead, and zinc) by its final chronic value (FCV[3]) to derive a sum of toxic units (SumTU).

For benthic invertebrates, it is believed that the narcosis mechanism determines the potency of sediment exposures to PAHs (USEPA 2003c, 2017a). Sediment was evaluated by summing the TOC-normalized

[3] The FCVs used in this calculation are those published in USEPA guidance (USEPA 2005b) and differ only slightly from the NJDEP chronic values listed in Table A-11. The only exception is lead, where the published FCV is 0.0081 ~~milligrams per liter (~~mg/L~~)~~ and the NJDEP chronic WQC is 0.024 mg/L.

concentrations divided by the benchmark for 34 PAHs (USEPA 2003c) to get the value $\sum ESBTU_{FCV}$. Porewater was evaluated by summing the concentrations divided by the FCV for 34 PAHs to derive a SumTU. When SumTU exceeds 1, there is a potential for toxicity to benthic invertebrates.

Table A-14 summarizes these evaluations. Under equilibrium partitioning theory, no toxicity is expected in any of the SQT sediment samples due to divalent metals concentrations, since all $\sum SEM\text{-}AVS/f_{OC}$ values are less than 130. Table A-14 also shows the SumTU calculations for porewater metals concentrations. A sample with a SumTU greater than 1 has the potential to exhibit chronic toxicity (USEPA 2005b). Cadmium and nickel were not detected in any porewater samples, so detected SumTU concentrations are based on some combination of the detected concentrations of copper, lead, or zinc. The porewater SumTU exceeded 1 at six stations (Stations 139, 159, 163, 141, 147, and 150) (Table A-14), indicating the potential for chronic toxicity at these locations as a result of copper, lead, or zinc in the porewater.

For PAHs, the $\sum ESBTU_{FCV}$ exceeds 1 at seven stations (Table A-14), indicating the potential for toxicity due to PAHs. In porewater the PAH SumTU exceeds 1 at one station, 154, at which acute toxicity was observed. The other station for which acute toxicity was observed, 151, has a PAH SumTU very close to 1.

2.3.5 Sediment and Porewater Relationship

Bivariate plots of sediment versus porewater concentrations for selected chemicals are shown in Attachment A-4. For many of the chemicals, there was a strong association between concentrations in sediment and concentrations in porewater. For example, porewater concentrations are shown to increase with sediment concentrations for 2,3,7,8-TCDD, total PCBs and total DDx (Figure A-4-1 and A-4-2). The relationship is not as strong for mercury (Figure A-4-3). However, for PAHs, there is no relationship (Figure A-4-2), suggesting that other factors control PAH concentrations in porewater.

2.3.6 Scoring

Based on the results of the screening of sediment and porewater concentrations, each SQT station was assigned a score ranging from 0 (unimpacted) to 1 (most impacted). Sediment and porewater were scored separately and combined in an equally weighted manner to derive the final score (Table A-15).

If a sediment sample had a concentration for any chemical that exceeded the threshold value shown in Table A-8, it received a score of 0.5. If a sediment sample had a concentration for any compound that exceeded the median value shown in Table A-8 and/or its ERM quotient exceeded 1, it received a score of 1.

If a porewater sample had a concentration for any chemical that exceeded a chronic criteria or screening value shown in Table A-11, it received a score 0.5. If the SumTU of the sample exceeded 1, it received a score of 0.5, regardless of whether any individual chronic criteria were exceeded. If a porewater sample had a concentration for any chemical that exceeded an acute criteria or screening value shown in Table A-11, it received a score of 1.

Sediment and porewater scores were equally weighted to derive a final score for each station (Table A-16). For sediment concentrations, all stations received the maximum score of 1 due to exceedances of the median sediment screening values or ERM quotients exceeding 1. These sediment screening values provide little in the way of helping to distinguish among stations, most of which display little to no

evidence of impact based on toxicity and BIC composition. For porewater concentrations, exceedances of chronic criteria were observed at 19 stations, 5 of which also had an exceedance of acute WQC. The metals SumTU exceeded 1 at six stations. The PAH SumTU exceeded 1 at one station, 154, where acute toxicity was also observed. No porewater exceedances were observed at 12 stations, 11 of which received an SQT score of 0 because neither SumTU exceeded 1. The combined sediment and porewater SQT scores ranged from 0.5 to 1.

2.4 Results

Table A-17 presents the combined final scores for each station based on each category of the triad. Consistent with the LPRSA BERA (Windward 2019) and as agreed upon with USEPA, the final scores were assigned impact categories as follows: no impact = <0.75; low impact = >0.75 and <1.5; medium impact = >1.5 and < 2.25; and high impact = >2.25. Most SQT stations had scores indicating no impact (10) or low impact (18). Two stations (Stations 151 and 154) were classified as medium impact. While the BIC metrics at these stations did not indicate impact, acute toxicity and elevated chemistry were observed. In fact, these stations had the highest concentrations observed in SQT sediment and porewater samples for many chemicals (Table A-18). No stations were classified as high impact.

3 STATISTICAL EVALUATION OF SQT DATA

3.1 Bivariate Correlation

Spearman rank correlation, a non-parametric measure, was used to evaluate the strength of the association between chemical concentrations and effects measures (i.e., toxicity endpoints and BIC metrics). The Spearman rank correlation coefficient is computed with ranked data and describes the strength of the monotonic association between two variables. Correlation coefficients can range in value from -1 to 1 and the absolute value describes the strength of the association, with 0 indicating no association and 1 or -1 indicating the strongest association. Positive values indicate one variable increases with the other. Negative values indicate one variable decreases with the other. Correlation coefficients, like all statistics, are subject to indicating an association by chance where none exists, and the probability of this is a function of sample size. For a sample size of 30, Spearman correlation coefficients with an absolute value greater than 0.3 indicate a statistically significant association (α = 0.05). It is important to note that a significant association does indicate causation of observed effects, particularly when multiple correlated chemicals are present in the sediment.

3.1.1 Sediment

Table A-19 presents Spearman correlation coefficients for sediment concentrations and effects measures. All toxicity endpoints and BIC metrics (except density), as presented in this table, are expected to decrease (rather than increase) as a result of elevated chemical concentrations in which case the association would result in a negative value for Spearman correlation. All significant and negative coefficients are ~~bolded~~bold in the table (i.e., all values less than -0.3). Acute toxicity, as expressed in the 10-day survival endpoint, was observed at only two of the 30 SQT stations (Stations 151 and 154). As shown in Table A-18, many chemicals were elevated in the sediment samples from those stations, resulting in the association of 10-day survival with many chemicals in this table, including dioxins and furans, PCBs, pesticides, PAHs, and metals. Only Aroclor PCBs were significantly negatively associated with 28-day mortality. Growth was negatively associated with multiple chemicals, including dioxins and furans, PCBs, pesticides, PAHs, metals, and TOC. Reproduction was negatively associated with PCBs, pesticides, PAHs, mercury, and TOC. The number of taxa was most strongly associated with percent fines and metals but was also associated with dioxins and furans, and PCBs. Indices of diversity, evenness, and dominance are not apparently negatively associated with sediment chemistry. The NY/NJ B-IBI is negatively associated with PCBs, PAHs, and methyl mercury.

3.1.2 Porewater

Table A-20 presents Spearman correlation coefficients for porewater concentrations and effects measures. All toxicity endpoints and BIC metrics (except density), as presented in this table, are expected to decrease (rather than increase) as a result of elevated chemical concentrations in which case the association would result in a negative value for Spearman correlation. All significant and negative coefficients are ~~bolded~~bold in the table (i.e., all values less than -0.3). Acute toxicity, as expressed in the 10-day survival endpoint, was observed at only two of the 30 SQT stations (Stations 151 and 154). As

shown in Table A-18, many chemicals were elevated in the porewater samples from those stations, resulting in the association of 10-day survival with many chemicals in this table, including dioxins and furans, PCBs, pesticides, and PAHs. With the exception of mercury and a few other metals (e.g., barium, chromium, iron, manganese), the FOD in porewater was low for metals; however, arsenic (Station 151 only), barium, and mercury were detected at their highest concentrations in porewater from the acutely toxic stations. At Station 151, arsenic exceeded the acute WQC. The 28-day survival endpoint was negatively associated with multiple chemicals, including dioxins and furans, PCBs, pesticides, PAHs, arsenic, and sulfide. Growth was negatively associated with multiple chemicals, including dioxins and furans, PCBs, pesticides, and PAHs. Reproduction was negatively associated with multiple chemicals, including dioxins and furans, PCBs, pesticides, PAHs, selenium, and sulfide. Conclusions regarding correlations with arsenic and selenium need to be interpreted with caution due to the lower FOD. Benthic metrics did not show any consistent association with classes of chemicals in porewater. For example, the number of taxa was negatively associated only with chromium, a single dioxin congener (1,2,3,7,8,9-Hexachlorodibenzo-*p*-dioxin), and a few PCB congeners.

3.1.3 Summary

Many chemicals, including dioxins and furans, PCBs, pesticides, PAHs, and metals, are significantly and negatively associated with toxicity endpoints in both sediment and porewater. While association does not indicate causation, there is evidence that chemical stressors may be responsible for the observed toxicity. The evidence is less clear for BIC metrics. For example, while several sediment chemical concentrations are negatively correlated with the number of taxa, this association is not seen for porewater. One explanation for this finding is that variability in BIC metrics is not a function of chemical concentrations and is mostly associated with a diverse community in which localized variability is a function of habitat. This is discussed further in the multivariate analysis sections below.

Based on the results of the SQT evaluation and the correlation analysis, a list of chemicals of potential concern to benthic invertebrates is provided in Table A-21. Chemicals on this list had exceedances of sediment and/or water quality guidelines/criteria and also were correlated with toxicity and/or BIC metrics for either sediment or porewater. Figures A-7a through A-7f show scatter plots of the 28-day survival endpoint versus sediment and porewater concentrations for a subset of the chemicals in Table A-21. In general, these graphs show that the samples with the lowest survival tend to have the highest concentrations in sediment and/or porewater of multiple chemicals, making it difficult to determine causation based on any single chemical. It is also evident that there is a lot of variation unexplained by chemical concentrations – a source of uncertainty in the interpretation of chronic endpoints. Because many individual chemicals are associated with effects measures, dose response modeling of individual chemicals would not be appropriate. Therefore, multivariate analyses were conducted to further evaluate the relationships.

3.2 Multivariate Analyses

3.2.1 Introduction

Because many chemicals are associated with effects measures, dose response modeling of individual chemicals would not be appropriate. Therefore, multivariate analyses were conducted to further evaluate

the relationships between chemicals and effects. One method of multivariate analysis is multivariate regression. In this method, multiple independent or predictor variables (e.g., chemical concentrations) can be evaluated with respect to their association with dependent or response variables (e.g., effects measures). However, when the predictor variables in a regression model are highly correlated, parameter estimates are unstable and have high standard errors – a problem referred to as collinearity or multicollinearity (Jackson 1991). A solution to this problem is to transform the predictor variables into principal components and use the principal components in subsequent modeling of the response variables (Jackson 1991). This transformation, called principal component analysis (PCA), was used to evaluate the relationships between sediment and porewater chemistry concentrations and response variables, because these chemicals are, in fact, highly correlated in the NBSA.

In PCA, a multivariate dataset is transformed using matrix algebra (i.e., orthogonal rotation) to create a new set of uncorrelated variables known as principal components (PCs), which are linear combinations of the original variables (i.e., chemical concentrations). PCs are sorted so that each, in turn, represents a smaller percentage of the variance within the dataset. These uncorrelated PCs, or some subset of them, can be used in subsequent linear modeling of response variables while avoiding the problem of multicollinearity. Another objective of PCA is to reduce the dimensionality of the dataset. When a multivariate dataset consists of many correlated variables, the structure of the dataset can be explained with fewer transformed variables (PCs). A common rule of thumb for PCA is that the first *n* principal components with eigenvalues greater than one are the most important and explain most of the variance of a multivariate dataset with correlated variables (Jackson 1991).

In addition to PCA, factor analysis was also used to evaluate the relationships between sediment and porewater chemistry concentrations and response variables. Factor analysis is a subsequent orthogonal rotation of the axes conducted to create new variables, called factors, that might be more easily interpretable than PCs. Varimax rotation was used, with the objective of having individual chemicals “load” (i.e., be strongly correlated with) a single factor and not load on any other factors. This is only achievable for multivariate datasets with simple structure, which is further described below.

PCA and factor analysis are subject to the assumptions of linear modeling, specifically that residuals are approximately normally distributed. Further, when variables that are measured in different scales and/or differ greatly in variance are combined in PCA, they are commonly rescaled by standard normalization (or other methods) so as not to give any one variable undue weight (Jackson 1991). These requirements were satisfied by conducting the PCA and factor rotation on a correlation matrix of natural logarithm transformed chemical concentrations. The PRINCOMP and FACTOR procedures in SAS© 9.4 were used to conduct the PCA and factor analysis.

3.2.2 Exploratory Principal Component Analysis Sediment Chemistry

The objective of dose response modeling of the SQT data is to be able to predict where toxicity to the BIC might be expected to occur in the NBSA. Therefore, a PCA model was developed using the entire Phase III (2014 to 2016) surface sediment dataset (n=254). Prior to conducting the PCA for sediment, exploratory analyses were conducted to determine a subset of chemicals or summed totals that could be used to represent large chemical groups and to simplify the analysis by reducing the variance introduced by chemicals that are unlikely to be causing the observed toxicity. This section describes the exploratory

analyses and how a final subset of 21 chemicals was chosen for the PCA sediment chemistry model of the NBSA.

The multivariate structure within each class of chemicals was examined using PCA. In PCA, when the input variables are highly correlated due to a single factor (i.e., some unmeasured or unmeasurable process), they will largely be explained by the first principal component (PC1) with subsequent PCs explaining either only minor other factors or just random variation. That property of PCA was used to aid in selecting either indicator chemicals or chemical summations to use in the final PCA. In multivariate analyses, all variables must have valid values. If any sample has missing values in a multivariate analysis, the entire sample is excluded. For samples with non-detect data, one-half the detection limit was substituted for non-detects. Chemicals with a FOD of less than 80% were excluded from the exploratory analyses due to the potential for adding additional unexplained variability.

A PCA of 16 dioxin and furan concentrations in sediment was conducted (Figure A-8a; left graph). Concentrations of 1,2,3,7,8,9-hexachlorodibenzofuran were excluded from the PCA due to a low FOD (61%). PC1 explains 91% of the variance of the 16 congeners indicative of the high correlation among them. The second principal component (PC2) explains 4.4% of the variance and, while it appears to represent a minor distinction among occurrences of dioxin and furan congeners, they are still highly correlated to each other. Based on this analysis, 2,3,7,8-tetrachlorodibenzo-*p*-dioxin (TCDD) in sediment was chosen to be carried forward in the PCA to represent all dioxin and furan congeners. Correlation coefficients for all other congeners with 2,3,7,8-TCDD exceeded 0.8.

A PCA was conducted of PCB congeners, or co-elutions, detected at least 80% of the time in sediment (Figure A-8a; right graph). PC1 explains 86% of the variance of the 100 remaining PCB congeners/co-elutions. PC2 explains 5.2% of the variance and appears to distinguish differences in the degree of chlorination among congeners (perhaps a weathering effect). However, since PC1 explains most of the variance of the PCB congeners, total PCB congeners (a single value) was chosen to be carried forward in the PCA to represent all PCB congeners.

A PCA of 32 PAH compounds in sediment was conducted (Figure A-8b; left graph). Six PAHs (C1-naphthalene, C4-chrysene, and parent and alkylated fluorenes) were excluded due to a FOD of less than 80%. PC1 explains 85% of the variance of the 32 remaining PAHs. PC2 and the third principal component (PC3) (not shown) explain 5.6% and 4.7% of the variance, respectively, and distinguish differences in molecular weight and alkylation among the individual PAHs (Figure A-8b; left graph). This distinction suggests that high molecular weight (HMW) PAHs, for example, are more highly correlated with each other than with other PAHs. The same goes for the low molecular weight (LMW) PAHs and the alkylated PAHs. This is not surprising given that pyrogenic and petrogenic sources of PAHs in the environment have unique PAH compositions. Because of these observations, three PAH summations were used to represent PAHs in the final PCA – total HMW PAHs, total LMW PAHs, and total alkylated PAHs. Total HMW and LMW PAHs refer to parent compounds only and have been previously defined (see Section 4 of the LPRSA BERA; Windward 2019). Total alkylated PAHs is the sum of the detected concentrations of the 18 alkylated PAHs measured in Phase III sediment samples.

A PCA of 20 metals in sediment was conducted (Figure A-8b; right graph). PC1 explains 78% of the variance and represents overall metal concentrations in sediments, which are all correlated to some extent. PC2 explains 10% of the variance and is differentiating metals, such as aluminum and iron, which are commonly found in fine-grained sediments, from others (e.g., mercury), which are found at elevated

concentrations due to industrial sources in the NBSA. Most metals that are positively correlated with PC2 are also highly correlated with aluminum. This includes the following metals and their correlation coefficients (ρ) with aluminum: beryllium (0.89), cobalt (0.89), iron (0.93), manganese (0.86), thallium (0.84), titanium (0.90), and vanadium (0.95). Their strong association with aluminium indicates that the presence of these metals is most likely associated with the rock-forming minerals commonly found in fine-grained sediments (nickel is the exception and is less correlated with aluminum: $\rho = 0.77$). Metals that are negatively correlated with PC2 (e.g., arsenic, chromium, mercury), while correlated with aluminum to some degree ($0.5 < \rho < 0.7$), are found at concentrations elevated above what might be associated with rock-forming minerals and are in exceedance of available sediment quality guidelines (Table A-8). Therefore, these metals, along with nickel, were included in the final sediment chemistry PCA model.

Other chemicals included in the final PCA model were based on detection frequency and the potential for toxicity to benthic organisms as follows:

- All 34 individual petroleum hydrocarbons were detected in surface sediments at frequencies ranging from 6 to 97% of samples. Although there were no sediment screening guidelines available for total petroleum hydrocarbons, they are potentially toxic to marine invertebrates (Hobbs et al. 2018) and were, therefore, included in the PCA. Total petroleum hydrocarbons, a summation of the concentrations of all 34 individual petroleum hydrocarbons, was used to represent this class of chemicals.
- Because of the differences in toxicity, no attempt was made to combine pesticides of different classes. Based on the observations previously described, such as detection frequency, exceedances of screening criteria, and correlation to effects measures, a subset of pesticides were selected for the PCA model. Total (2,4' and 4,4') DDx was used to represent all DDx isomers. Among chlordane chemicals, alpha- and gamma-chlordane were detected most frequently and often at values greater than screening guidelines; therefore, total alpha + gamma chlordane was used to represent chlordane chemicals. Hexachlorobenzene and dieldrin, both of which have high rates of detection in sediment and were fully detected in porewater, were also included in the PCA.
- No semivolatile organic compounds were included in the final PCA model. None were significantly correlated with toxicity and the low detection frequency (which would require substitution for non-detects because no values can be missing in the PCA) would add uncertainty to the model.

A preliminary PCA was conducted with all sediment samples. It became apparent that two samples collected during the 2016 Phase III sediment sampling program in the area of the confined disposal facility (CDF) between the port channels, were multivariate outliers (NB03SED-CHM349 and NB03SED-CHM354). Chemical concentrations in the uncharacteristically sandy sediment samples from these locations were very low or non-detect, adding additional variance to the model. These samples, which are not representative of the NBSA and are more likely representative of capping material on the CDF, were excluded from further multivariate analysis. In addition, another 2016 Phase III sediment sample near the confined disposal facility (NB03SED-CHM339) was excluded because of missing pesticide data. Concentrations of other chemicals were low at this location. This left a total of 251 Phase III surface sediment samples included in the sediment chemistry PCA. Detection frequencies were high among the remaining samples and 21 chemicals (>95%). One-half the detection limit was substituted for any remaining non-detected results before the data transformations were applied.

3.2.3 Principal Component Analysis Sediment Chemistry

Table A-22 provides the correlation matrix of natural logarithm transformed chemical concentrations of the 21 chemicals selected for the PCA. This table also shows the eigenvalues of the first six PCs. PC1 explains 66.7% of the variance, and the first three PCs explain a total of 83% of the variance of the sediment chemistry dataset (Table A-22). All remaining PCs have eigenvalues less than 1. All 21 chemicals are correlated to PC1 (Figure A-9). PC2, explaining 9.4% of the variance, is correlated with chlordane, dieldrin, 2,3,7,8-TCDD, and total PCBs. PC3, explaining 6.5% of the variance, is correlated with PAHs. While PC1 explains most of the variance in overall chemical concentrations, PC2 and PC3 explain some relative differences in concentrations among the chemical groups. Because PCs are new variables that are linear combinations of the original variables (i.e., chemical concentrations), the concentration data can be used to find the value, usually referred to as the score, of each PC for each sediment sample. Because PC1 represents overall chemistry concentrations, it is useful to view in map form to see patterns of elevated chemistry. Figure A-10 presents the scores on PC1 of the Phase III surface sediment samples in the NBSA. Lower (more negative) PC1 scores indicate lower concentrations, while scores near 0 indicate median concentrations. Greater (more positive) scores indicate higher chemical concentrations, and it will be shown that PC1 is a good predictor of toxicity.

3.2.4 Factor Analysis Sediment Chemistry

In addition to PCA, factor analysis was conducted on the sediment chemistry data to determine if individual chemicals, or chemical classes, could be assigned to factors that could potentially explain the observed toxicity. Based on the outcome of the PCA, it was hypothesized that three factors would explain the variance of the multivariate dataset. Varimax rotation resulted in the factor loadings that are shown on Figure A-11. The objective of varimax rotation is to have each chemical “load” onto (i.e., be strongly correlated with) one factor and no other. This was not fully achievable for the NBSA sediment chemistry dataset; however, some chemicals loaded more strongly on individual factors. Metals were most strongly correlated to Factor 1 (Figure A-11). Pesticides, PCBs, and 2,3,7,8-TCDD were most strongly correlated to Factor 2. However, the pesticides – total DDx and hexachlorobenzene – also had significant loading on Factor 1 with metals. PAHs and total petroleum hydrocarbons were most strongly correlated to Factor 3. However, alkylated PAHs and total petroleum hydrocarbons also had significant loading on Factor 1. While simple structure is not achieved by the factor model, the factors can still be tested in dose response models to examine whether certain groups of chemicals appear to have a greater association with effects.

3.2.5 Dose Response Modeling Sediment

The objective of conducting the PCA was to create a set of variables, or PCs, that represent sediment chemical concentrations (but are uncorrelated and, therefore, are not subject to the problem of multicollinearity) to be used in a quantitative effects assessment. Linear modeling methods were used to evaluate the relationship between predictor variables (e.g., PCs, grain size) and response variables (e.g., 10-day mortality, number of taxa). Due to the nature of the response variables, two types of models were used: 1) linear/multivariate regression models – appropriate for continuous response variables (e.g., growth) and 2) binary response models – the appropriate model form for variables with only two possible outcomes (e.g., survival/mortality).

Linear regression models have the following form:

$$R = \alpha + \beta_i x_i + \varepsilon \qquad \textbf{Equation 3-1}$$

where:

R = response (e.g., growth, number of taxa)

α = intercept parameter

β_i = slope parameter of the *ith* predictor variable

x_i = *ith* predictor variable (e.g., PC1), and

ε = random error.

Competing linear regression models were compared using the adjusted coefficient of determination (R^2). The adjusted R^2 gives an indication of the ability of the model to explain the variance of the response data while penalizing the introduction of additional spurious predictors to the model. Models were fitted using the general linear model procedure (PROC GLM) in SAS© 9.4.

Survival data from the 10- and 28-day *L. plumulosus* bioassays were evaluated using linear binary response models. Because of the natural, or threshold, mortality expected during the test, as measured in the control, it was more convenient to model mortality rather than survival: (%Mortality = 100-%Survival). These models have the following form:

$$\pi = C + (1 - C)F(\alpha + \beta_i x_i) + \varepsilon \qquad \textbf{Equation 3-2}$$

where:

π = probability of mortality

C = threshold mortality (i.e., mean laboratory control mortality)

F = a cumulative probability distribution function

α = intercept parameter

β_i = slope parameter of the *ith* predictor variable

x_i = *ith* predictor variable (e.g., PC1), and

ε = random error.

The cumulative distribution function (*F*) maps the range of the linear function $(-\infty,\infty)$ onto the range of probability (0,1). Three different probability distributions were tested for each subset of model predictors: the logistic distribution, the standard normal distribution (called a probit model), and the Gompertz distribution (also called the extreme value model). Competing binary response models were compared using Akaike's Information Criteria (AIC). Given a set of candidate models, the preferred model is the one with the minimum AIC value (Akaike 1981). Binary response models were fitted using the probit procedure (PROC PROBIT) in SAS© 9.4. The PROBIT procedure calculates maximum likelihood estimates of regression parameters for binary response data, including probit, logit, and extreme value (i.e., Gompertz link) regression models. If convergence of the maximum likelihood estimates is attained, a Type III chi-square test statistic is computed for each effect parameter specified in the model, testing whether there is any contribution from any of the levels of the effect parameter. This statistic is computed as a quadratic form in the appropriate parameter estimates by using the corresponding submatrix of the

asymptotic covariance matrix estimate. The AIC is calculated from the log likelihood ln(L) for each model as follows:

$$AIC = -2 \ln(L) + 2p \qquad \textit{Equation 3-3}$$

where:

ln(L) = log likelihood

p = the number of parameters in the model, including the intercept

Thus, AIC rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of predictors in the model.

For each response variable, the following predictors were tested: PC1, PC2, PC3, TOC, and percent fines. For BIC metrics, shear stress[4] was also tested. Predictor variables were added to each model in a stepwise (i.e. step up and step down) fashion. In addition to the adjusted R^2 and the AIC, the Type III marginal test statistic, computed for each parameter specified in the model, was examined to determine whether there was any contribution from the added predictor variable. Non-significant predictors (α = 0.05) were removed from the final models.

3.2.5.1 PCA Models

Before any model fitting was conducted, plots of the PCs and toxicity endpoints were constructed. Figure A-12 shows 10-day survival versus the three PCs, and it is apparent that there is a strong association between PC1 and 10-day survival. Table A-23 presents the details of the models tested for toxicity endpoints and the sediment chemistry PCA. For the modeled endpoint, 10-day mortality, PC1 was the only significant predictor. Among the three model forms tested, the probit model had the lowest AIC. The fitted model is depicted on Figure A-13 and shows the probability of mortality increasing with the value of PC1, which is strongly correlated to overall sediment chemical concentrations (Figure A-9). Based on this model, the predicted probability of mortality in all Phase III sediments was calculated. Predicted mortality was transformed back to survival (%Survival = 100~~-%~~ %Mortality), and samples for which survival is predicted to be less than 80% and/or 60% of the control survival were identified on a map (Figure A-14). These locations are labeled with their abbreviated location identifiers (i.e., station identifier). Based on the model, acute toxicity is expected in samples with a PC1 score of 5 or greater for which there are a total of 21 (8.2% of all Phase III samples) in the NBSA – three in the north geographic area and 18 in the south geographic area.

Figure A-15 shows 28-day survival versus the three PCs. This survival endpoint exhibits more variability than 10-day survival and the associations are less obvious. For this endpoint, the best model was a Gompertz model with both PC1 (overall chemistry) and PC2 (pesticides, dioxin, PCBs) as predictors (Table A-23). It had the lowest AIC of any tested model for which all parameters were significant according to the Type III statistic. This model is depicted on Figure A-16 (PC2 not shown). It was observed that 28-day mortality was much more variable than 10-day mortality and, therefore, the predictability of the model is less certain. However, based on the observed results, the model correctly

[4] Shear stress is defined as the daily maximum grain stress in dynes per square centimeter predicted by the hydrodynamic model (1995 to 2013) of the NBSA in the model grid cell in which the SQT sample was located.

predicts toxicity (or lack thereof) 83% (25 out of 30) of the time (Figure A-16; inset table). If toxicity is defined simply as 28-day observed survival significantly different from control, then six stations would be classified as toxic (Table A-5). Four of these stations (Stations 143, 151, 154, and 160) are also predicted to be toxic by the model. Two of the six were predicted to be non-toxic. Of these two, Station 141, with observed survival of 67%, was statistically significantly different from control but was not less than 80% of control. The model predicted survival for this station is 73% (90% of control), which, by definition, is not toxic. Therefore, the model agrees with the observed result for this station. The other observed toxic station for which the model predicts to be non-toxic, Station 158, shows a significant lack of agreement with the model (Figure A-16). Observed survival was 39% (i.e., 61% mortality), while the model-predicted survival was 81%. This station does not have elevated chemical concentrations for most chemicals except PAHs. In fact, total PAHs in porewater was greatest at Station 158 (Figure A-7d), which could be contributing to the observed toxicity that is not explained by sediment concentrations. In addition, there was considerable variability in replicate survival at this station, with replicate survival ranging from 0 to 75% (Attachment A-2). The model predicts three stations to be toxic that were not observed to be toxic. For two of these (Stations 138 and 149), survival was within the confidence limits predicted by the model and, had they been statistically significant compared to control, they would have been classified as toxic with an observed survival of 46% and 63%, respectively. The third station, Station 161, had a very high observed survival (96%) but had a model-predicted survival of 59%. In summary, the model correctly predicts toxicity 83% of the time. Of the five stations for which observed and predicted toxicity were not in agreement, only two stations (158 and 161) exhibited observed survival outside the range of the model confidence limits. Along with station 158, station 142 also showed significant lack of agreement with the model (Figure A-16). Although it was not statistically significantly different from control, and therefore judged non-toxic, observed survival was 53% (i.e., 47% mortality) at station 142. Model-predicted survival is 81% for this station and therefore the station was correctly predicted to be non-toxic despite the difference in observed and predicted survival. Similar to station 158, there was considerable replicate variability at this location with replicate survival ranging from 0 to 95% (Attachment A-2).

Similarly, as was done for the 10-day model, 28-day predicted probability of mortality in all Phase III sediments was calculated and transformed back to survival. Samples for which 28-day survival is predicted to be less than 80% and/or 60% of the control survival were identified on a map (Figure A-17). Chronic toxicity is predicted in 45 (17.7%) of the Phase III samples – 11 in the north geographic area, mostly in the same area where acute toxicity was predicted; five spread out in the central geographic area; and the remaining 29 in the south geographic area, mostly in the same areas where acute toxicity was predicted.

Growth versus the sediment chemistry PCs are shown on Figure A-18. Two stations (Stations 151 and 160) exhibited 100% mortality during the 28-day test. These stations are depicted with a value of 0 on the graphs for illustrative purposes, but growth and reproduction were not measured at these stations and were treated as missing values in dose response modeling. PC1 was the only significant predictor of growth (Table A-23). The fitted model is depicted on Figure A-19. Mean growth in the laboratory control sample was 0.045 mg/organism/day. Only one SQT sample met or exceeded this value (Figure A-18; Table A-5). The SQT sample with the lowest overall chemical concentrations, from Station 155, had an observed growth of 0.035 mg/organism/day. As previously discussed, the sublethal endpoints are subject to considerably more variability that is unrelated to chemical stressors. Kennedy et al. (2009) found significant variability in control sediments for growth with a CV of 56% as well as significant variability in

an interlaboratory comparison of growth. Eickhoff et al. (2014) found considerable variability in mean growth in five control sediments from the same estuary, ranging from 0.025 to 0.064 mg/organism/day. The only test acceptability criterion for control growth is that it is measurable (USEPA 2001), setting a low bar for this endpoint. Growth was measurable in all but two of the SQT samples that exhibited 100% mortality at 28 days. It is unknown and unknowable what growth rate could be achieved in the NBSA sediments in the absence of chemical stressors. However, based on the results from Station 155, it is more likely to be in the range of 0.035 mg/organism/day or less. At the median PC1 score of 0, the regression model predicted a mean growth of 0.03 mg/organism/day – a 35% reduction compared to the control but within the range of growth observed in control sediments by Eickhoff et al. (2014). At a PC1 score of 5, the threshold for prediction of acute toxicity, the regression model predicted a mean growth of 0.02 mg/organism/day – a reduction of 57% from control growth.

The reproductive endpoint versus the sediment chemistry PCs are shown on Figure A-20. The reproductive endpoint was not significantly associated with any of the predictors tested (Table A-23; Figure A-21), which is not surprising given the variability in this endpoint. Reproduction in the laboratory control (mean = 1.97 young per surviving adult) had significant replicate variability, ranging from 0.33 to 4 young per surviving adult (CV = 73%). Among the SQT samples, reproduction ranged from 0 to 1.93 (mean = 0.68 young per surviving adult) (Table A-5). As previously discussed, the reproductive endpoint is subject to considerable variability unrelated to chemical stressors. Eickhoff et al. (2014) observed mean reproduction in five control samples from the same estuary, ranging from 0.8 to 8.4 young per surviving adult with an average replicate CV of 79%. Kennedy et al. (2009) found significant variability in control sediment for reproduction with a CV of 63%, as well as significant variability in an interlaboratory comparison of reproduction. Reproduction is also known to be highly dependent on grain size (McDonald et al. 2010).

Scatter plots of each benthic metric versus PC1 are shown on Figures A-22a through A-22c. In addition, the percent tolerant and percent sensitive of total density metrics, which are part of the B-IBI calculation, are also shown versus PC1 (Figure A-22d). The detailed results of model testing for the BIC metrics are shown in Table A-24. None of the PCs was found to be a significant negative predictor for BIC metrics other than density. Density was negatively associated with PC1 (Figure A-23a). Inferences about density versus chemical concentrations are subject to uncertainty, because chemical stressors can result in both increases and decreases in this metric (Weisberg et al. 1998). The number of taxa was most associated with grain size (Table A-24; Figure A-23b), with an increasing percentage of silt and clay resulting in a decline in the number of taxa. PC1 is positively associated with evenness, a spurious finding since elevated chemistry is not expected to result in an increase in evenness. Percent fines has a significant positive association with evenness (Figure A-23c).

Table A-25 provides the PC scores and predicted mortality for 251 Phase II samples evaluated and depicted on Figures A-14 and A-17.

3.2.5.2 Factor Models

The factors described in Section 3.2.4 were also tested using the same dose response modeling methods described in Section 3.2.5 (Table A-26) in an effort to determine if any distinctions could be made as to which chemicals or chemical groups might be causing toxicity. For 10-day mortality, both Factors 1 (metals, DDx, alkylated PAHs) and 3 (PAHs, total petroleum hydrocarbons) were significantly and

positively associated with mortality (Table A-26; Figure A-24). For 28-day mortality, both Factors 1 (metals, DDx, alkylated PAHs) and 2 (pesticides, PCBs, and 2,3,7,8-TCDD) were significantly and positively associated with mortality (Table A-26; Figure A-25). These findings suggest that multiple chemicals contribute to the observed toxicity. No factor models were significant for growth or reproduction (Table A-26). These findings are consistent with the findings in the PC models, where PC1, representing elevated chemistry for all chemicals, was a significant predictor of both 10- and 28-day toxicity, and PC2 was also significant for 28-day mortality.

3.2.5.3 Summary

The PCA, factor analysis, and dose response modeling provide evidence that chemical concentrations in sediments are associated with toxicity to benthic invertebrates. The PCA-based dose response models can be used to reasonably predict where toxicity has the potential to occur in the NBSA, as shown on Figures A-14 and A-17. The factor-based dose response models confirm that multiple chemical mixtures are associated with the observed toxicity. Major findings include:

- The PCA of sediment chemistry concentrations indicates that most of the variance is explained by the first three PCs with PC1 explaining the most variance because it is representative of overall concentration and most of the chemicals are correlated with each other (Figure A-13).
- The 10-day toxicity testing resulted in only two toxic stations with low survival while all other stations had very high survival. These two stations also had the highest concentrations of many chemicals in sediment and porewater (Table A-18) resulting in ~~a strong~~an association between acute toxicity and PC1. The uncertainty with respect to estimating a threshold for acute toxicity is discussed in Section 4.1.
- The 28-day survival results were subject to considerably more variability and the relationship to chemistry, while significant, is not as strong as it is for the 10-day tests. Both PC1 and PC2 are associated with 28-day survival (Figure A-16). Two stations, 142 and 158, that showed significant departure from the model (Figure A-16) also had significant within replicate variability for this endpoint.

3.2.6 Multivariate Analyses for Porewater

PCA was also conducted using the porewater chemistry data. To be consistent with the sediment chemistry PCA, where possible, the same subset of chemicals was used in the porewater PCA:

- Porewater concentrations of 2,3,7,8-TCDD, detected in all porewater samples, were used to represent all dioxin and furan congeners.
- Most PCB congeners had high rates of detection in porewater, with 111 congeners/co-elutions having a 100% detection rate. Total PCB congeners was used to represent all PCB congeners in the PCA.
- Individual PAH concentrations are represented by two totals – total PAHs and total alkylated PAHs (both 100% detection rates). Total LMW PAHs had a FOD rate of only 50% in porewater and were, therefore, not evaluated separately from HMW PAHs as was done for sediment.

- Except for mercury, only 29 samples were analyzed for metals due to a broken dialysis bag for the sample from Station 155. Of the remaining samples, most metals had a very low FOD: antimony (3%), arsenic (31%), cadmium (0%), copper (17%), lead (21%), nickel (0%), silver (0%), and zinc (10%). Therefore, only mercury, which was detected in 29 of 30 samples, was included in the porewater PCA model. One-half the detection limit was substituted for any non-detects.
- Total (2,4' and 4,4') DDx was used to represent all DDx isomers and was fully detected in porewater. Among chlordane chemicals, alpha- and gamma-chlordane were fully detected, with concentrations as much as 10 times greater than other chlordane chemicals; therefore, total alpha + gamma chlordane was used to represent chlordane chemicals. Hexachlorobenzene and dieldrin, both fully detected in porewater, were also included in the PCA.
- Semivolatile organic compounds and petroleum hydrocarbons were not analyzed in porewater.

Table A-27 provides the correlation matrix of the natural logarithm transformed porewater concentrations for the nine chemicals and the eigenvalues. PC1 explains 51% of the variance, and the first three PCs explain a total of 80% of the variance of the sediment chemistry dataset (Table A-27; Figure A-26). All remaining PCs have eigenvalues less than 1. All but mercury and total alpha + gamma chlordane are correlated to PC1. PC2, explaining 19% of the variance, is correlated with mercury and total alpha + gamma chlordane. PC3, explaining 10% of the variance, is correlated with mercury and PAHs. While PC1 explains most of the variance in overall chemical concentrations, PC2 and PC3 explain some relative differences in concentrations among the chemical groups.

Based on the PCA, three factors were evaluated for porewater (Figure A-27) using varimax rotation, with the objective of having chemicals strongly load (i.e., be correlated with) a single factor and no other factors. However, this was not achieved for the porewater chemistry dataset. Total PCB congeners, 2,3,7,8-TCDD, hexachlorobenzene, and dieldrin are most strongly correlated with Factor 1 (45% of the variance). PAHs are most strongly correlated to Factor 2 (35% of the variance), and mercury is most strongly correlated with Factor 3 (20% of the variance). Total DDx and alpha + gamma chlordane are correlated with all factors; although, in opposite directions in Factors 2 and 3. Total DDx is positively correlated to Factor 2 and negatively correlated to Factor 3, while the opposite is observed for alpha + gamma chlordane.

3.2.6.1 Porewater Dose Response Models

Both porewater PCs (Table A-28) and porewater factors (Table A-29) were evaluated in dose response modeling of porewater chemistry and toxicity. The modeling methods were the same as those described in Section 3.2.5 for sediment. Only PCs and factors were tested as predictors for toxicity. PC1 (and no other PCs) is significantly and positively associated with 10-day mortality (Table A-28; Figure A-28), 28-day mortality (Table A-28; Figure A-29), and growth (Table A-28; Figure A-30). Reproduction was not significantly associated with any PCs (Table A-28; Figure A-31).

Both Factors 1 (pesticides, PCBs, and 2,3,7,8-TCDD) and 2 (PAHs and DDx) were significantly and positively associated with 10-day mortality (Table A-29; Figures A-32a and A-32b); however, combining them in one model resulted in unstable parameter estimates. This indicates that both factors may be giving similar or redundant information, which makes sense given that both of the two acutely toxic stations, Stations 151 and 154, have relatively high scores on both factors. Therefore, the chemicals (or

some subset of them) associated with PC1 and both Factors 1 and 2 (i.e., all except mercury and total alpha + gamma chlordane) are potentially contributing to the observed acute toxicity. Factors 1, 2, and 3 (mercury) were significantly and positively associated with 28-day mortality (Table A-29; Figure A-33), again indicating a multiple chemical effect. Factor 2 (PAHs and DDx) is significantly and negatively correlated with growth (Table A-29; Figure A-34a). When added to the model, Factor 1 improved the adjusted R^2 but was not significant (p=0.06). Despite lack of significance, the improvement in the R^2 value suggests that Factor 1 may also be negatively predictive of growth after controlling for Factor 2 (Table A-29; Figure A-34b).

Porewater PCs were also evaluated as predictors of BIC metrics (Table A-30) and were not found to be significant predictors of any BIC metrics.

3.3 Community Ordination

Benthic metrics are useful indicators of BIC health, but some information is lost in the analysis of metrics only. Therefore, community ordination was also evaluated with respect to sediment chemistry. Ordination refers to a variety of techniques used to arrange benthic samples in relation to one or more coordinate axes and to each other to provide information about their ecological similarities. Community abundance data collected over environmental gradients are typically non-linear and are better suited to non-linear methods of analysis, such as non-metric multidimensional scaling (NMDS) of a Bray-Curtis Dissimilarity Matrix (Ludwig and Reynolds 1988). Like PCA, NMDS is a method used to reduce the dimensionality of the community data, the result of which is that each BIC sample can be scored in the new smaller coordinate space. These scores can be plotted on graphs where samples near each other are judged more similar than those further away on the graph. Spatial patterns may be discernible if the BIC varies based on geographic or geomorphic location. The resulting dimensions can be regressed against other variables, such as individual chemicals, PCs, or factors, to evaluate the effects of stressors on changes in community structure.

Figures A-35a through A35e illustrate the results of the NMDS evaluation. Goodness of fit criteria indicate that the community structure can be represented in two dimensions (Figures A-35a and A-35b). The space between samples can be thought of as distances that increase with dissimilarity. For example, Stations 148 and 160 have no species in common and exhibit the maximum distance between any two stations on the graph. In this type of analysis, the dimensions are not always easily interpretable; however, dimension 2 is correlated with the number of taxa. There are no obvious geographical trends (Figure A-35a). Other spatial trends, based on geomorphic unit or USEPA Subunit, were also not apparent (Figure A-35c). Linear egression modeling was conducted to determine if the sediment chemistry PC1 was a predictor of community structure (Figure A-35d). There is no significant relationship between PC1 and dimension 1 (p=0.26) nor between PC1 and dimension 2 (p=0.38). It might appear from the graph (Figure A-35d; right graph) that a relationship between PC1 and dimension 2 would be more apparent if the two toxic stations (Stations 151 and 154) were removed from the analysis; however, the p-value of this model is still not significant (p=0.09). Other predictors for BIC dimensions were tested: PC2, PC3, TOC, and percent fines (Table A-31). The only finding was that percent fines (Figure A-35e) is a significant predictor for dimension 2 (p=0.004; R^2 = 0.26), which is not surprising given that percent fines was also the best predictor for the number of taxa.

In summary, there is no evidence of geographic spatial patterns among the benthic community structure and no apparent relationship to sediment chemistry. Sediment grain size is the best predictor of patterns in community structure.

3.4 Historical Data Evaluation

In 1993 to 1994, 1998, 2003, 2008, and 2013, USEPA, under REMAP, collected co-located sediment, toxicity (i.e., 10-day *Ampelisca abdita* survival bioassay), and BIC samples at 140 random locations in the area referred to as the Newark Bay sub-basin, which, in addition to the Bay, included stations in the Arthur Kill, Lower Passaic River, and Hackensack River (Adams 2016; USEPA 2003a). A total of 28 locations were sampled within the Newark Bay sub-basin each year of the study, and between 11 and 18 stations (sampling year dependent) were located within the NBSA boundary. See Figure 2-2 of the main BERA document for a map of the locations. Attachment 5 provides the chemistry (Table A-5-1) toxicity (Table A-5-2) and BIC data (Table A-5-3).

Table A-32 presents a Spearman correlation matrix of sediment concentrations versus survival and BIC metrics for the NBSA samples in each REMAP study. Negative correlations indicate that a chemical concentration is associated with a decrease in the metric. All metrics except proportion tolerant species are expected to decrease with degradation. The proportion tolerant species are expected to increase with degradation. Similar relationships are apparent in the REMAP data as was seen in the 2015 SQT data. For example, metals (including simultaneously extracted metals), pesticides (e.g., total DDx, chlordanes, hexachlorobenzene) and PCBs are negatively correlated (positively for proportion tolerant) with amphipod survival and BIC metrics. PAHs are negatively correlated with survival in the 1998 study but in general, associations with PAHs in sediment are not strong. This was similar to the 2015 SQT data where sediment correlations with PAHs were not strong; however, there was a stronger association with PAHs in the SQT porewater samples and toxicity and BIC metrics. The dioxin congener, 2,3,7,8-TCDD, was negatively correlated with survival in the 2003 study, but not the two other studies for which dioxins and furans were measured.

In earlier studies conducted by NOAA in two phases within the Hudson-Raritan estuary, which includes Newark Bay, 10-day *Ampelisca abdita* survival was also found to be correlated with several different chemical classes. In Phase 1 (1991), survival was significantly, negatively, and strongly correlated with PAHs. This relationship was attributed to high toxicity and PAH concentrations in samples from the East River, New York. In Phase 2 (1993) which was focused on Newark Bay and its tributaries, survival was significantly negatively correlated with PCBs, pesticides (total DDx, chlordanes, dieldrin, hexachlorobenzene), dioxins, and metals, a finding consistent with the findings for the 2015 SQT samples.

In general, these historical studies have demonstrated a decline in the probability of toxicity in the NBSA over time. Amphipod survival for the Phase 2 NOAA study and the REMAP studies are depicted on Figure A-36. In the NOAA study, 30 of 34 samples (88%) collected in the NBSA had survival less than 80% of control. In subsequent studies, the percentage of toxic samples (defined as less than 80% of control survival) was 22% (REMAP 1993/94), 54% (REMAP 1998), 9% (REMAP 2003), 9% (REMAP 2008), and 0% (REMAP 2013). In the current study, 10-day *L. plumulosus* survival less than 80% of control was seen in 2 out of 30 samples (7%). The decline in toxicity coincides with improved indicators of the health of the BIC in these studies as described in Section 2.3.1 of the main BERA document.

3.5 Summary

The statistical analysis and quantitative effects assessment demonstrated the following conclusions.

Acute toxicity is significantly correlated with chemical concentrations in sediment and porewater. The PCA evaluations demonstrated that measures of overall chemical concentrations, as represented by the first PC in both the sediment and porewater PCA, are significantly correlated with 10-day *L. plumulosus* mortality and, based on this relationship, areas where acute toxicity might occur can be reasonably predicted based on the Phase III sediment chemistry PCA model (Figure A-14). The factor analyses demonstrated that toxicity is associated with a mixture of multiple chemicals: metals, pesticides, PCBs, PAHs, and dioxins/furans. No single chemical or chemical group in the evaluation was shown to be unrelated to toxicity. It is important to note that, because of the high correlation among chemical concentrations, these findings cannot prove that all individual chemicals evaluated are causing toxicity. Acute toxicity was only observed at two of the 30 SQT stations in the NBSA. Those two stations (151 and 154) happened to have the highest concentrations in both sediment and porewater of most chemicals, including those chemicals incorporated in the multivariate analysis (Table A-18).

Chronic toxicity is also significantly correlated with chemical concentrations in sediment and porewater, as demonstrated by the PCA models. And similarly factor analysis demonstrated that toxicity is associated with a mixture of multiple chemicals: metals, pesticides, PCBs, PAHs, and dioxins/furans. Based on models of 28-day mortality, areas where chronic toxicity could potentially occur can be reasonably predicted based on the Phase III sediment chemistry (Figure A-17).

For chronic sublethal endpoints, growth, but not reproduction, is significantly negatively correlated with chemical concentrations; although there is significant variability in the sublethal responses.

With the exception of density, none of the BIC metrics or community ordination measures were significantly correlated with chemical concentrations. PC1 is a significant predictor of density ($p=0.026$; $R^2 = 0.17$); however, because density can either increase or decrease as a result of degradation (Weisberg et al.1998), this relationship is not evidence of chemical impact to the BIC.

Similar patterns of correlation between toxicity and sediment chemistry are seen in the historical SQT data from the NBSA.

4 SUMMARY OF SQT LINES OF EVIDENCE

Table A-33 summarizes the SQT LOEs based on the individual evaluations of BIC, toxicity, and sediment and porewater data and the statistical evaluations. The dose response modeling and comparisons to Jamaica Bay reference indicate that there is no apparent impact to the BIC due to chemical stressors. Multivariate analyses indicate that acute and chronic toxicity is associated with chemical mixtures in sediment and porewater. Toxicity occurred in samples with elevated concentrations of multiple chemicals: dioxins/furans; PCBs; pesticides (particularly total DDx, dieldrin, chlordane, and hexachlorobenzene); PAHs (both parent an alkylated); petroleum hydrocarbons; and most inorganics (e.g., arsenic, cadmium, chromium, copper, lead, mercury, and zinc). Areas where the dose response models predict potential toxicity are of limited extent in the NBSA and are primarily located in the southwestern subtidal flat and in a smaller area in the north in the vicinity of historically disturbed sediments.

4.1 Uncertainties

There are several uncertainties associated with this SQT assessment related to the following issues: weighting and scoring in the SQT evaluation, estimation of porewater concentrations, prediction of acute toxicity, variability in chronic toxicity endpoints, reference comparisons and the representativeness of the 30 SQT samples. These are each discussed in the following paragraphs.

The scoring in the SQT evaluation used equal weights for each category of the triad: BIC, toxicity, and sediment chemistry. Alternative scoring methods have been proposed in the literature that give less weight to data with more uncertainty, in particular, to the sediment chemistry category, because it is generally based on a comparison to screening values (i.e., T20/T50 and ERL/ERM) that have not been demonstrated to accurately predict toxicity as they were derived from field sediments with chemical mixtures. In this assessment, the sediment chemistry score is based on an equal weighting of sediment and porewater chemistry. While the sediment chemistry portion of the score is based on a comparison to screening values, the porewater chemistry portion of the score is based on exceedances of promulgated WQC, which may be a better indicator of exposure and potential impact for benthic communities. Therefore, to assess this uncertainty, the scoring was recalculated to give porewater chemistry the full weight of the sediment chemistry category score (i.e., sediment chemistry is ignored in the scoring). The result of the scoring change is that seven of the low impact stations become no impact stations for a total of 17 no impact stations. The two stations categorized as medium impact (Stations 151 and 154) remain so. The remaining 11 stations keep their low impact classification (Table A-6-1 of Attachment A-6). This alternative scoring is perhaps a better characterization of risk given the paucity of observed impacts to BIC and the low levels of observed acute toxicity. Another alternative scoring scenario was evaluated that assigned a score of 1 to any survival value less than 80% of control survival (Table A-6-2 of Attachment A-6), regardless of statistical significance. This was evaluated due to the fact that variability in the replicate data could result in MDDs that would not identify samples as toxic when survival is less than 80% of control (see Attachment A-2 for estimates of MDDs for each endpoint). The outcome of this scoring is 9 no impact stations and 19 low impact stations. The medium impact stations maintain their categorization.

While it is expected that porewater concentrations would provide better information about exposure, it is important to note the uncertainty in the measurement of porewater chemistry. Concentrations of organic compounds in porewater were obtained by ex-situ exposure of PE and POM passive samplers to sediments in a laboratory-controlled setting. PE passive samplers loaded with PRCs were used to determine the freely dissolved porewater concentrations of PCBs, PAHs, and pesticides. POM passive samplers (without PRCs) were used to measure freely dissolved concentrations of dioxins and furans. Following exposure to sediment, the PE and POM samplers were transferred to the analytical laboratories where the target organic compounds were extracted in solvent (acetone/hexane [1:1 by volume]) and the concentrations in final extracts were measured. Porewater concentrations were estimated using the laboratory-reported concentrations in the PE/POM extracts as described by Ghosh (2014). A simple mass balance calculation is used to convert extract concentrations to concentrations within the PE/POM. The porewater concentration is estimated as the ratio of the PE/POM concentration and a literature-derived K_{PS}. The PRCs were used to assess the extent of equilibrium achieved during the period that the samplers were in contact with the sediment. Some of the strongly hydrophobic compounds, in particular, the higher chlorinated PCB congeners, did not achieve equilibrium during the period of contact, and a mathematical correction was performed to calculate the true equilibrium concentrations (Fernandez et al. 2009). Uncertainty in the porewater concentrations arises from the fact that the source of the K_{PS} values are regression relationships derived in other experiments with these samplers (Ghosh 2014; Choi et al. 2013; Fernandez 2012; Cornelissen et al. 2008). Another source of uncertainty for the dioxin/furan concentrations is that PRCs were not used in the POM samplers and, therefore, no equilibrium corrections were made. There is a possibility that the dioxin/furan concentrations are biased low; however, if non-equilibrium occurred, it is more of a concern for the higher-chlorinated dioxin/furan congeners that have much lower toxicity equivalent factors compared to, for example, 2,3,7,8-TCDD. For PCBs, non-equilibrium occurred infrequently and only among the higher-chlorinated congeners (i.e., hexa-, hepta-, octa-, nona-, and deca-chlorinated congeners).

The uncertainty of K_{PS} values used in this SQT assessment is further described in the USEPA's passive sampling user's manual (USEPA/SERDP/ESTCP 2017). Estimated values of K_{PS} are expected to be accurate within 0.3 log units (i.e., a factor of 2) even for highly hydrophobic compounds. The resulting error is similar to other environmental partition coefficients. Therefore, porewater concentrations could be as much as two times greater or lower than the calculated values presented here. As an uncertainty analysis, porewater concentrations for organic chemicals were doubled and then compared to the acute and chronic values presented in Table A-11. The chronic exceedance counts increased for PCBs and DDx beyond that shown in Table A-12. For PCBs, doubling the concentrations resulted in five additional samples exceeding the chronic criteria for a total of eight out of 30. For total DDx, doubling the concentrations resulted in 15 additional samples exceeding the chronic criteria for a total of 29 out of 30. The doubling also resulted in one sample (i.e., station 154) exceeding the chronic freshwater SQuiRT value for 2,3,7,8-TCDD (Table A-11) since the estimated value changed of 9.35 picograms per liter (pg/L) exceeds 10 pg/L when doubled. No additional acute exceedances resulted from the doubling. The PAH SumTU values shown in Table A-14 can also be doubled to evaluate the uncertainty. Doubling results in one additional sample (i.e., station 151) with a PAH SumTU that exceeds 1.

Modeling of acute toxicity indicated a strong association between sediment and porewater chemistry and 10-day survival. Because there were only two stations that exhibited toxicity, there is uncertainty that the relationship is real or could have happened by chance. However, a review of the historical data shows a

clear relationship between 10-day *A. abdita* survival and sediment chemistry in the NBSA (see Section 3.4); therefore, it is very likely that the same relationship exists with *L. plumulosus* as both organisms have demonstrated ability to identify toxic sediments (Schlekat et al. 1995). There is also uncertainty in the model fitting because all but two stations had very high survival. The PC1 scores of the toxic stations, 151 and 154, are 7.3 and 5.6, respectively. The next highest PC1 score for the non-toxic samples is 3.52. There are four non-toxic stations with PC1 scores that exceed 3; therefore, it is clear that the threshold for acute toxicity is somewhere between 3.5 and 5.6. The sediment model predicts toxicity at PC1 scores exceeding 5 and identified 21 Phase III locations as potentially acutely toxic. If the true PC1 score threshold for toxicity were lower, for example, 4, then an additional 13 Phase III locations would also be classified as potentially acutely toxic for a total of 34, 13.4% of all samples. If the true PC1 score threshold for toxicity was 3.6, then an additional 16 Phase III locations would also be classified as potentially acutely toxic for a total of 37, 14.6% of Phase III samples.

There is uncertainty associated with the results of toxicity tests. Endpoint response to sediments is not necessarily indicative of contaminant effects (Kennedy et al. 2009). A reduction in endpoint values (i.e., survival, growth, number of young) compared to control could be caused by other factors, including inherent random variability that could affect both the tested samples and/or the control samples. The 10-day acute toxicity tests have control acceptability criteria that require >90% survival; therefore, control variability is tightly controlled in these tests. Nonetheless, clean reference samples have been shown to exhibit 10-day *L. plumulosus* survival less than control (Kennedy et al. 2009).

The 28-day tests are subject to more variability. In an interlaboratory comparison, Kennedy et al. (2009) found greater inconsistency among chronic endpoints as compared to acute testing results. For 28-day survival, Kennedy et al. (2009) and Eickhoff et al. (2014) found fairly low variability among laboratory control samples: within replicates, CVs ranged from 2 to 11%. However, among test sediments, toxicity classifications differed among duplicate testing results (Kennedy et al. 2009). For example, a sediment sample from Newark Bay that was tested three different times exhibited 28-day survival significantly less than control (i.e., toxic) in the first test and comparable to control in subsequent tests (i.e., non-toxic). A clean reference sample exhibited 28-day survival of 60%, which would result in classification as toxic.

The sublethal endpoints are subject to considerably more variability in both control and test sediments. There are no control acceptability requirements that control variability in these endpoints. Eickhoff et al. (2014) subjected five control sediments to 28-day tests with *L. plumulosus* and found that the growth and reproduction endpoints were more variable both among control samples and within batch replicates. Mean growth in control samples ranged from 0.7 to 1.8 milligrams per organism (0.025 to 0.064 mg/organism/day), and the mean CV among replicates was 30%. Reproduction had a 10-fold range within control sediments (0.8 to 8.4 juveniles per surviving adult), and the mean CV among replicates was 79%. Kennedy et al. (2009) found significant variability in control sediment for growth (CV = 56%) and reproduction (CV = 63%) as well as significant variability in an interlaboratory comparison of the growth and reproduction endpoints. *L. plumulosus* reproduction is known to be highly dependent on grain size (McDonald et al. 2010). This variability leads to uncertainty in the interpretation of these chronic tests.

Variability in the sublethal endpoints was also demonstrated in split sample testing conducted by the USEPA on a subset of three SQT samples (USEPA 2017b). Growth was found to be significantly different than control in samples from Stations 144, 156, and 157 (Table A-7). However, the split results indicated growth greater than control for these same sediments. This inconsistency is not unlike what has been

documented in the literature and confirms the inherent uncertainty in the interpretation of sublethal endpoints.

For the BIC comparisons to reference, there is uncertainty due to the difference in sampling depths, 2 cm for Jamaica Bay and 6 inches in the NBSA, which may cause differences in abundance and richness in the samples. Both programs incorporated three grab samples per location for taxa enumeration. In both programs, BIC samples were sieved with a 0.5 millimeter sieve.

There is uncertainty regarding the incorporation of the 2013 REMAP data from Jamaica Bay into the reference dataset. As noted in Section 2.3.1 of the BERA, there was a significant decline in the BIC metrics when 2013 REMAP data are compared to 2008 REMAP data (and prior) in the NBSA. This same pattern was also observed in Jamaica Bay. This decline affected only BIC measures, as there was no change in the incidence of toxicity. The mean chemical concentrations presented in Attachment A-1 (Table A-1-5) are similar between the two programs. It is not known why this apparent decline in BIC occurred; however, given it was observed in both estuaries, it is due to either of two possibilities: 1) a change in sampling methodology (unlikely, due to REMAP using similar methods since 1993); or 2) changes caused by a regional event. In fact, the 2013 samples were collected after a significant meteorological event called Hurricane Sandy occurred in October 2012. It is possible the storm affected the BIC. In the NBSA, these effects, if real, were short-lived, because the 2015 dataset indicated a full recovery. There are currently no available data to infer whether the BIC has rebounded in Jamaica Bay since 2013. To evaluate this uncertainty, the 2013 REMAP data were removed from the Jamaica Bay reference dataset and the reference envelope was recalculated. The recalculated values can be found in Table A-1-1. The change in the reference envelope had no impact on the scoring of the NBSA BIC data. The 5th and 95th percentiles did not change significantly enough to make a difference. All NBSA BIC metrics are within the reference envelope either way except for abundance at station 148 as noted when the entire data set is used. The removal of the 2013 data had no impact on the evaluation. There were no additional exceedances of the reference envelope nor changes in SQT scoring, and the hypothesis testing results were similar in that there was no indication the NBSA had lower mean BIC values than Jamaica Bay.

There is uncertainty regarding whether the 30 SQT stations are representative of the larger Phase III surface sediment data sets. This was evaluated in two ways, first, by comparing summary statistics of selected chemicals for the 30 SQT stations and the remaining Phase III samples (Table A-3-3 of Attachment A-3). Mean concentrations are similar for the SQT samples and the remaining samples. For example, the mean 2,3,7,8-TCDD concentration in SQT samples is 91.3 nanograms per kilogram (ng/kg) while for the remaining Phase III samples, it is 86.6 ng/kg. Other chemical concentrations compare similarly. The mean SQT versus remaining samples are as follows respectively, total PCB congeners (588 micrograms per kilogram [µg/kg], 465 µg/kg), total 2,4~~4~~4’- and 4,4~~-4~~4’-DDx (98.7 µg/kg, 135 µg/kg), total PAHs (13,000 µg/kg, 12,000 µg/kg), and mercury (1.92 milligrams per kilogram [mg/kg~~,~~], 2.4 mg/kg). The mean and median concentrations for other chemicals are also similar as well~~ as TOC and~~. The mean percent fines for the SQT samples (60.3%) is higher than for remaining Phase III data (55.5%) and mean TOC for the SQT data (40,100 mg/kg) is also higher than for the remaining data (21,400 mg/kg). This could be due in part to the finer grained samples in the SQT data set. In addition, an extreme outlier for TOC (151,000 mg/kg) occurs in the SQT data set (Station 154). With the outlier excluded, the mean TOC in the SQT data set is 36,300 mg/kg. Higher TOC may decrease the bioavailability of organic compounds. The data sets were also compared visually by plotting the principal component scores color-coded by

sampling program (Figure A-9b). The spatial pattern shown on the graph indicates that the SQT samples have a similar range and chemical composition as the remaining sediment samples. These evaluations indicate that the SQT samples are representative of the full Phase III data set and the NBSA surface sediments.

In comments to the draft BERA, the USEPA provided analyses of alternative methods of conducting PCA and multivariate analyses. These analyses are found in Attachment A-7. The analyses conducted by USEPA support the PCA approach for multivariate analyses and dose response modeling presented in this BERA as discussed in the following paragraphs.

The USEPA conducted PCA on data transformed by the “row sum of 1” transformation such that each variable represents a mass fraction of a contaminant rather than an absolute concentration. These analyses are presented in Figures 1 through 6 of Attachment A-7. Based on these analyses, the USEPA concluded “the PCA on mass fraction using the “on Correlations” method is useful to identify sources or to fingerprint and seems to reveal some structure in the data and relationships among variables, but not to identify risks associated with exposures to contaminants”.

The USEPA conducted PCA on the SQT data only (Figures 7 and 8 of Attachment A-7) and incorporated some additional pesticides (i.e., BHC and other chlordane summations) and TOC to the analysis. The results are similar to those presented in this BERA demonstrating that the PCA model is relatively insensitive to different data and variable combinations.

The USEPA conducted PCA on other subsets of the Phase III data set and also incorporated historical data (Figures 9 and 10 of Attachment A-7). The results show similar relationships as were found in the PCA presented in this BERA; specifically, the first principal component is correlated with both 10- and 28-day amphipod survival. For example, compare Figure 10 of Attachment 7 to Figure A-12 (left panel) and Figure A-15 (left panel). Again, this demonstrates that the PCA model is relatively insensitive to different data and variable combinations.

Finally, the USEPA used a simpler method than PCA for a multivariate analysis. Recognizing that PC1, in describing the largest variance, is somewhat similar to an average of the standardized chemicals, ~~this~~the sum of the standardized concentrations of 18 chemicals (i.e., the 18 chemicals listed in Attachment A-7 for historical data) was plotted against survival (Figure 11 of Attachment A-7) demonstrating a similar relationship as was shown in the BERA for PC1. For example, compare Figure 11 of Attachment A-7 to Figure A-12 (left panel) and Figure A-15 (left panel). Figure 12 of Attachment A-7 shows areas where the sum of standardized concentrations exceeds 5. These areas are similar to the areas for which toxicity was predicted by the PCA model (Figures A-14 and A-17).

Figure 11 of Attachment A-7 also shows the historical REMAP data collected in 2003 and 2013. A similar analysis of the earlier REMAP data collected in 1993/94 and 1998 is shown in Figure A-37 indicating that there has been a relationship between sediment chemistry and 10-day amphipod survival in the NBSA historically.

5 REFERENCES

Adams, D. 2016. Overview of EPA sediment monitoring activities. Presentation at the New Jersey Water Council Meeting. September 26.

Akaike, H. 1981. Likelihood of a Model and Information Criteria. Journal of Econometrics 16:3-14.

American Public Health Association, American Water Works Association, and Water Pollution Control Federation. 1980. Standard for Methods for the Examination of Water and Wastewater. Fifteenth Edition. Washington, D.C.

ASTM International. 2008. E-1267-03. Standard Guide for Measuring the Toxicity of Sediment-Associated Contaminants with Marine and Estuarine Amphipods.

Buchman, M.F. 2008. NOAA Screening Quick Reference Tables, NOAA OR&R Report 08-1. National Oceanic and Atmospheric Administration, Office of Response and Restoration Division. Seattle, Washington. 34 pp.

Choi, Y., Y-M Cho, and R.G. Luthy. 2013. Polyethylene−water partitioning coefficients for parent- and alkylated-polycyclic aromatic hydrocarbons and polychlorinated biphenyls. *Environmental Science & Technology*, 47, 6943-6950.

Cornelissen, G., I.T. Cousins, K. Wiberg, M. Tysklind, H. Holmström, and D. Broman. 2008. Black carbon-dominated PCDD/Fs sorption to soils at a former wood impregnation site. *Chemosphere* 72(10): 1455-1461.

Eickhoff, C., M. Grey, J. Pickard, S. Reimer, and D. MacDonald. 2014. The co-dependency of growth and reproduction: a closer look at the 28 day *Leptocheirus plumulosus* survival, growth and reproduction test. A presentation at the Society of Environmental Toxicology and Chemistry Annual Meeting, Vancouver, British Columbia, Canada. November.

Fernandez, L.A., C.F. Harvey, and P.M. Gschwend. 2009. Using performance reference compounds in polyethylene passive samplers to deduce sediment porewater concentrations for numerous target chemicals. *Environmental Science & Technology* 43:8888-8894.

Fernandez, L.A., W. Lao, K.A. Maruya, C. White, and R.M. Burgess. 2012. Passive sampling to measure baseline dissolved persistent organic pollutant concentrations in the water column of the Palos Verdes Shelf Superfund site. *Environmental Science & Technology* 46 (21): 11937-11947.

Fernandez, L.A., W. Lao, K.A. Maruya, and R.M. Burgess. 2014. Calculating the diffusive flux of persistent organic pollutants between sediments and the water column on the Palos Verdes Shelf Superfund Site using polymeric passive samplers. *Environmental Science & Technology*, 48 (7): 3925-3934.

Ghosh, U., S.K. Driscoll, R.M. Burgess, M.T.O. Jonker, D. Reible, F. Gobas, Y. Choi, S.E. Apitz, K.A. Maruya, W.R. Gala, M. Mortimer, and C. Beegan. 2014. Passive sampling methods for contaminated sediments: practical guidance for selection, calibration, and implementation. Integrated Environmental Assessment & Management, DOI: 10.1002/ieam. 1507.

GSH. 2017a. Newark Bay Study Area Sediment Quality Triad and Porewater Field Report. Revision 1. September.

GSH. 2017b. Newark Bay Study Area Sediment Quality Triad and Porewater Data Report, Revision 2. September.

Hobbs, W., J. Weakland, and C. Eickhoff, 2018. Environmental Effects-Based Concentrations for Total Petroleum Hydrocarbons (TPH): Toxicity in Marine Water and Freshwater. Washington State Department of Ecology. Olympia, Washington. Publication No. 18-03-002.

Jackson, E.J. 1991. A User's Guide to Principal Components. John Wiley & Sons, Inc.

Kennedy, A.J., J.A. Steevens, G.R. Lotufo, and T.S. Bridges. 2009. A comparison of acute and chronic toxicity methods for marine sediments. *Mar Environ Res* 68(3):118-127.

Long, E.R., D.D. MacDonald, S.L. Smith, and F.D. Calder. 1995. Incidence of adverse biological effects within ranges of chemical concentrations in marine and estuarine sediments. *Environ Manage* 19:81-97.

Long, E.R., C.G. Ingersoll and D.D. MacDonald. 2006. Calculation and uses of mean sediment quality guideline quotients: A critical review. Environmental Science & Technology 40:1726-1736.

Ludwig, J.A. and J.F. Reynolds. 1988. Statistical Ecology: A Primer on Methods and Computing. John Wiley and Sons, New York.

McDonald, B.G., C.A. McPherson, R. DeWynter, P. Kickham, and C. Brown. 2010. Incorporation of 28-d *Leptocheirus plumulosus* toxicity data in a sediment weight-of-evidence framework. *Ecotox Environ Safe* 73:51-55.

NJDEP. 2016. New Jersey Administrative Code 7:9B Surface Water Quality Standards. October 17.

NOAA, 1995. Magnitude and Extent of Sediment Toxicity in the Hudson-Raritan Estuary. Technical Memorandum NOS ORCA 88. National Status and Trends Program for Marine Environmental Quality. National Ocean Service. Office of Ocean Resources Conservation and Assessment. Coastal Monitoring and Bioeffects Assessment Division. Silver Spring, Maryland.

Pielou, E.C. 1966. The measurement of diversity in different types of biological collections. *J Theor Biol* 13:131-144.

Schlekat, C.E., K.J. Scott, R.C. Swartz, B. Albrecht, L. Antrim, K. Doe, S. Douglas, J.A. Ferretti, D.J. Hansen, D.W. Moore, C. Mueller, and A. Tang. 1995. Interlaboratory comparison of a 10-day sediment toxicity test method using *Ampelisca abdita*, *Eohaustorius estuarius* and *Leptocheirus plumulosus*. *Environmental Toxicology and Chemistry* 14:2163-2174.

Shannon, C.E. 1948. A mathematical theory of communication. *Bell Systems Technical Journal* 27(July and October):379-423 and 623-656.

Swartz, R.C., D.W. Schultz, G.R. Ditsworth, W.A. DeBen, and F.A. Cole. 1985. Sediment toxicity, contamination, and macrobenthic communities near a large sewage outfall. In: Boyle T.T., ed, Validation and predictability of laboratory methods for assessing the fate and effects of contaminants in aquatic ecosystems. STP 865. American Society for Testing and Materials. Philadelphia, Pennsylvania, pp 152-175.

Tierra. 2015. Newark Bay Study Area Sediment Quality Triad and Porewater Sampling and Analysis, Quality Assurance Project Plan, Revision 2. August.

USEPA. 1994. Methods for Assessing the Toxicity of Sediment-Associated Contaminants with Estuarine and Marine Amphipods. EPA/600/R-94/025.

USEPA. 2001. Method for Assessing the Chronic Toxicity of Marine and Estuarine Sediment-Associated Contaminants with the Amphipod *Leptocheirus plumulosus*. Office of Research and Development Western Ecology Division. U.S. Environmental Protection Agency. Newport, OR. EPA 600/R-01/020.

USEPA. 2003a. Sediment Quality of the NY/NJ Harbor System: A 5-year Revisit (1993/1994 to 1998). U.S. Environmental Protection Agency, Region II. Edison, New Jersey. EPA/902-R-03-002.

USEPA. 2003b. Procedures for the Derivation of Equilibrium Partitioning Sediment Benchmarks (ESBs) for the Protection of Benthic Organisms: Endrin. U.S. Environmental Protection Agency, Office of Research and Development. Washington, D.C. EPA-600-R-02-009.

USEPA. 2003c. Procedures for the Derivation of Equilibrium Partitioning Sediment Benchmarks (ESBs) for the Protection of Benthic Organisms: PAH Mixtures. EPA-600-R-02-013. Office of Research and Development. Washington, DC, USA.

USEPA. 2005a. Predicting Toxicity to Amphipods from Sediment Chemistry. United States Environmental Protection Agency, National Center for Environmental Assessment. Washington, D.C. EPA/600/R-04/030.

USEPA. 2005b. Procedures for the Derivation of Equilibrium Partitioning Sediment Benchmarks (ESBs) for the Protection of Benthic Organisms: Metal Mixtures (Cadmium, Copper, Lead, Nickel, Silver and Zinc). U.S. Environmental Protection Agency, Office of Research and Development. Washington, D.C. EPA-600-R-02-011.

USEPA 2017a. Developing Sediment Remediation Goals at Superfund Sites Based on Pore Water for the Protection of Benthic Organisms from Direct Toxicity to Non-ionic Organic Contaminants. EPA/600/R. Office of Research and Development. Washington, DC, USA.

USEPA. 2017b. Correspondence. Final Oversight Report: 2014-2016 Softshell Clam, Blue Crab, and Fish Oversight; 2015 Sediment Quality Triad Oversight; Newark Bay Study Area RI/FS Oversight. Prepared by Louis Berger. June 30, 2017.

USEPA/SERDP/ESTCP. 2017. Laboratory, Field, and Analytical Procedures for Using Passive Sampling in the Evaluation of Contaminated Sediments: User's Manual. EPA/600/R-16/357. Office of Research and Development, Washington, DC. https://www.epa.gov/superfund/superfund-contaminated-sediments-guidance-documents-fact-sheets-and-policies.

USEPA and USACE. 1998. Evaluation of Material Proposed for Discharge to Waters of the US – Testing Manual (Inland Testing Manual). United States Environmental Protection Agency. Washington, D.C. EPA/823/B-98/004.

Weisberg, S.B., J.A. Ranasinghe, J.S. O'Connor, and D.A. Adams. 1998. A Benthic Index of Biotic Integrity (B-IBI) for the New York/New Jersey Harbor. Appendix C in Adams, D.A., O'Connor,

J.S., Weisberg, S.B. 1998. Sediment quality of the NY/NJ Harbor System. United States Environmental Protection Agency, Region II. Edison, New Jersey. EPA/902-R-98-001.

Windward. 2019. Lower Passaic River Study Area Baseline Ecological Risk Assessment, Final. June 17.

TABLES

FIGURES

APPENDIX A-1

Jamaica Bay Reference Data

APPENDIX A-2

Leptocheirus plumulosus Bioassay Data

APPENDIX A-3

Sediment and Porewater Data Screens

APPENDIX A-4

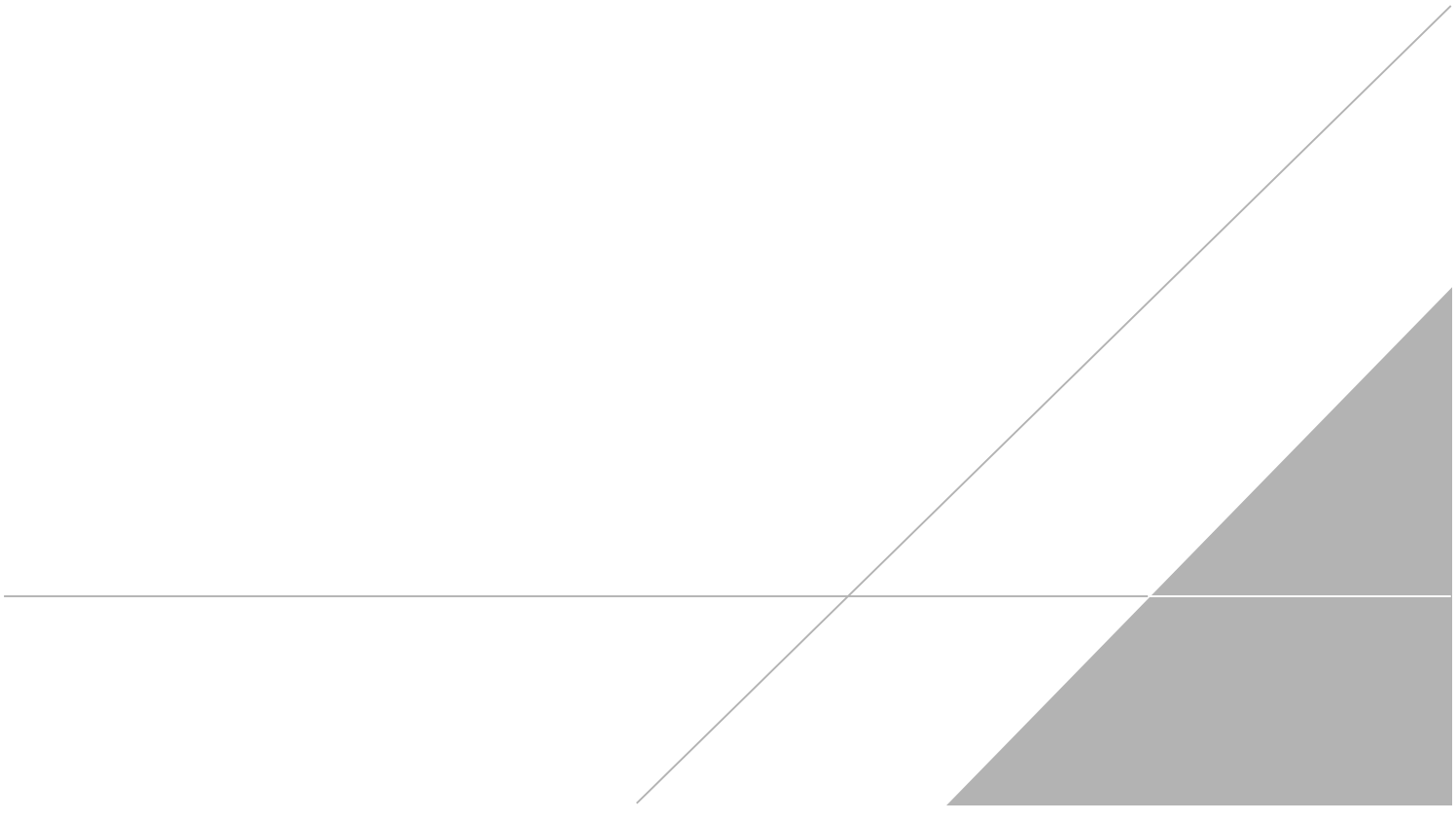
Porewater vs. Sediment Concentrations

APPENDIX A-5

Historical SQT Data for the NBSA

APPENDIX A-6

SQT Scoring Uncertainty Evaluations



APPENDIX A-7

Multivariate Analyses Conducted by USEPA